Department of Sociology
King's University College
Sociology 3306b: Section 570 Investigating the Social World: Quantitative Research

Computer assignment #3

Regression Analysis Assignment/ Regression Procedures for Final Paper

SOLELY FOR THOSE USING BINARY LOGISTIC REGRESSION!!
If you are using "linear OLS regression" for your final paper, consult the other assignment posted.

At this point in the term, you should have a pretty good sense as to which variables and what dataset you will be working with in your final research paper. You should also have a reasonable sense as to the principal hypotheses that you are putting to empirical test.
The purpose of the current assignment is twofold.

First, it will provide you with a simple strategy for dealing with sample weights when conducting multivariate analyses (sample weights relate to the sample design: I will talk further about them in class). Second, it will provide you with a strategy for conducting your final logistic regression analysis for your final paper. The results from this assignment can be directly plugged into your final paper.

Part A: Weighting

What do I mean by "unweighted" as opposed to "weighted" data?

The principle behind estimation in a probability sample such as the GSS, CCHS or NLSCY or the public use sample from the census (3%) is that each person in the sample "represents", besides himself or herself, several other persons not in the sample. For example, in a 2% simple random sample of the population, each person in the sample represents 50 persons in the population. The weighting phase is a step which calculates, for each record, what this number is (i.e., the number of individuals in the population represented by this record). The relevant weights appear on the respective datasets as variables. In the NLSCY microdata file the weight is called "AWTCW01". With the Census, the name of the relevant weight is "WEIGHT". In the CCHS, the relevant weight is called "WTS_M". In the GSS, the relevant varaible is called "WGHT_PER".

These should be considered when deriving meaningful estimates from the survey.

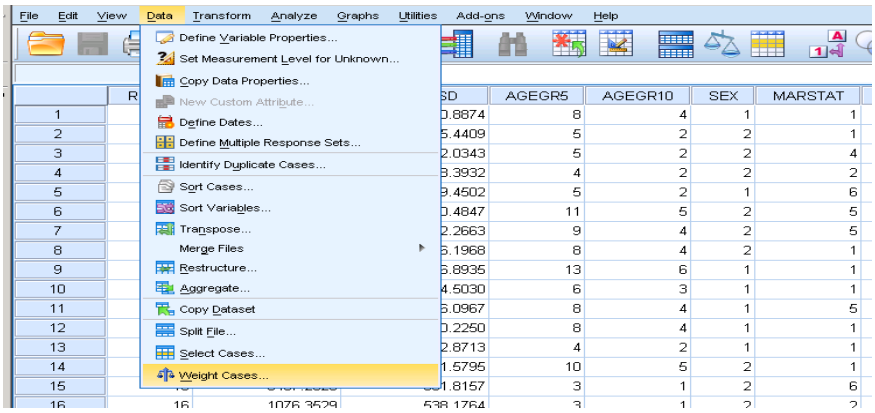
In a probability sample, the sample design itself determines weights which must be used to produce unbiased estimates of the population. Each record (i.e. each case in your sample) must be weighted by the inverse of the probability of selecting the person to whom the record refers. In the example of a 2% simple random sample, this probability would be .02 for each person and the records must be weighted by 1/.02=50. The importance of working with weights is obvious when you start to examine the complex character of many of Statistics Canada's sample designs. In other words, Statistics Canada works with very complex sampling designs whereby it is clear that not all individuals in the targeted population have an equal likelihood of being selected.

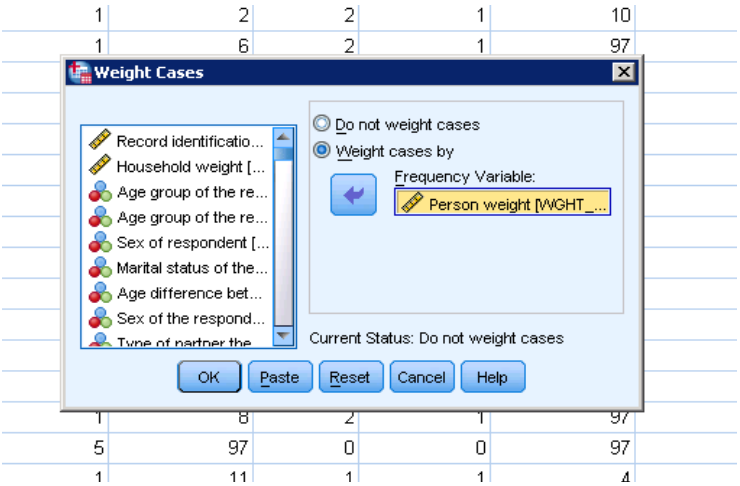
Therefore, in the absence of simple random sampling (whereby every case would have an equal weight), it is necessary to rely upon sample weights in obtaining descriptive statistics that are not biased by the sample design. Note: this is not much of an issue with the Census public use file as it is merely a simple random sample of the full census database and not a complex sample design (i.e. each person, with a few exceptions, represents roughly the same number persons in the full population).

Step 1.

Run a frequency distribution on "Province of Residence" from your dataset (use the dataset that you have selected for your term paper). P**rint-up the syntax and output file.** This gives you an "unweighted" distribution of your sample on province of residence. Use the TITLE procedure to properly document all steps in this assignment. No write up required here.

Step 2. Weight your data and run this frequency distribution again. Using SPSS it is easy to "weight" your dataset. Go to Data>Weight Cases, and then specify the weight appropriate to your dataset **(**AWTCW01, WEIGHT, WTS_M or WGHT_PER).





Paste this procedure into a syntax file, and then below it place the exact same frequency command as specified above (on province of residence).

When you run this syntax file, this should give you the "weighted" frequency distribution on province of residence. **Print up the respective syntax and output files (properly titled). Briefly comment on how it is different relative to the frequency distribution obtained in Step 1 above.** The weighted results are meant to be "more" representative of the Canadian population.

Do you now have a feel for what is meant by weighting your data? You take the sample, which might have a somewhat complex sampling design, and use the sample weights to make our statistics more representative of the larger population. For example, most of Stats Canada's samples are stratified by province. Typically they "over-sample" the smaller provinces (e.g. PEI, Newfoundland), and "under sample" the larger ones (Quebec or Ontario). This is done for purely statistical reasons, to obtain reasonable estimates for all provinces. It is easy to adjust the estimates with the appropriate sample weights after the fact. Consider the distribution across provinces in the weighted and un-weighted distributions. The "weighted" estimates are obviously more representative than the "non-weighted" estimates (PEI has a population of only about 150,000 persons or less than ½ of a percent of Canada's total: what do "un-weighted" estimates suggest in terms of percentage living in this province? (I will discuss further in class).

Step 3. Important point: In Step 2 you "weighted the data". If you are asked to work with the "unweighted data" again (as is the case below) you must first "deactivate" the weights (by going back to your weight command, and pasting into your syntax file the "do not use weights" command). Alternatively, if you close and re-open your datafile, your data will also no longer be "weighted".

A problem when running the weight feature on SPSS is that it unfortunately it can lead to *erroneous inferences with regard to statistical significance*. This is particularly the case with **regression procedures**. The reason for this is SPSS continues to treat your data like it were a sample, even though your calculations now involve the weighted numbers (for example, your sample might be only 19000 cases, but the weighted results make reference to tens of millions of Canadians). Remember that if you work with very large samples, then even small differences or small associations may be significant.

A simple way to deal with this problem is as follows: Obtain the "mean" of the sample weight corresponding to the data that you are using in your analysis **(**AWTCW01, WEIGHT, WTS_M or WGHT_PER). You can obtain the mean using the "Descriptives" command in SPSS (go to "Analyze">"Descriptive Statistics"> "Descriptives" and select your "Weight" variable with this procedure. The output should give you the mean. No need to provide a syntax file or output file with this step (merely record the "mean" on this weight variable for your own records).

Next, with this mean value, compute a new set of weights. The new weights should be equal to the initial weights divided by this mean score. For example, when working with the NLSCY, assume that you found using the Descriptives procedure a mean on AWTCW01 of 200.3342 (note: the mean for the NLSCY weight does not equal 200.3342, but I am merely using this number for demonstrative purposes). This would imply that on average each person in your sample represented about 200 persons. Your revised weights for your analysis could be calculated by using the following command:

Syntax for new weights: COMPUTE NWEIGHT= AWTCW01 / 200.3342

Create a new syntax file with this command, and run your syntax file. After doing so, you have created a new set of weights (let us refer to NWEIGHT in the above example as our new "weight variable").

With this new weight variable, you must re-weight your data. This is very easy to do. Again, use the Data>Weight cases procedure (selecting this new weight variable that you've just created which should be at the very end of your list of variables) and paste this in your syntax file.

Alternatively, you could alternatively merely place under your COMPUTE command:

WEIGHT BY NWEIGHT.

When you run this weighting procedure, you have just reweighted your data again. The results should be "unbiased" by sample design but now the TOTAL reported should be the same size as your initial sample size (allowing for meaningful inference tests).

To assure that you have done this properly, again run a "Frequency distribution on the same province of residence variable, yet this time with the "re-weighted" data. Place the frequency command in the same syntax file, yet make sure it is below the aforementioned COMPUTE and WEIGHT commands. Run this procedure and comment.

Note: If you have done this properly, you should have a distribution from Step 3 that has the same 'Total" as your initial "un-weighted sample" in Step 1 yet have the same "percentage distribution" as your weighted result obtained in Step 2.

Print up your syntax and output files from Part 3, properly titled, and include a brief comment on results.

Part B. Target population

If applicable, using SPSS, carefully select the subsample that you will be focusing on in your final paper (example, all children aged 4-11 or all adults aged 18+). As in previous assignments, use the "SELECT CASES" procedure in delineating the sample for your analysis. Again, this relates to the "target population" of your analysis.

If your variables that you select for your research are only applicable to certain subsamples within the GSS, CHHS, NLSCY or Census public use file (e.g. an age appropriate behavioral scale) then you must explicitly take this into consideration when you select your subsample for your multiple regression. Also note that it doesn't make sense to select subsamples if there is no strong theoretical or technical reason to do so. **You need not hand anything in for Step 1, merely report the dataset and your subsample (if applicable) or state that you are working with the full sample.**

PART C. Model Building

From now on, we shall be using "weighted" data in running **ALL REMAINING REGRESSIONS in this assignment.** This will allow for "unbiased results" that take into consideration your sample design while also allowing for more meaningful significance tests.

Although there are various strategies possible when building models (stepwise regression; forward selection procedures, among other largely automated procedures), the best of all possible strategies is let your "theory" guide you in terms of introducing variables into multiple regression. This is what I am recommending for the current assignment and for your final paper, whether you are working with linear or logistic regression. This is also why you are currently conducting and writing up a "literature review", i.e. you are trying to situate your empirical analysis in terms of a broader research literature. Remember: if your results are non-significant, don't worry… that's okay. Merely report it in your final paper.

Working with Weighted Data:
So far in Assignments 1 and 2, you have worked exclusively with "unweighted" data. For the current exercise, you will also conduct several regressions using "weighted" data. **Again, the results from Part C can be used directly in your final paper.**

For all the following steps, your data MUST be properly weighted. If you use the full sample (no subsample) for your analysis, you can merely use the weights that you created in Part A above (i.e. make sure that all the syntax files that you use in Part C are weighted by the weight that you created in Part A (note: not the initial weights, but the new weights you created). If you are selecting a sub-sample for your analysis, it is necessary to quickly repeat what you did in Part A, yet this time exclusively for your selected sub-sample:

I) select your subsample
II) obtain the mean of the initial weight (using descriptives, again using either AWTCW01, WEIGHT, WTS_M or WGHT_PER).
III) create a new weight using this mean and the initial weight (using compute as in Part A)
IV) make sure that all of your syntax files (Steps 1 – 3 below) used from here on in are weighted by this new weight

Step 1-3 below must all be properly documented (TITLE) and weighted (WEIGHT).
Merely place your appropriate WEIGHT command directly below your TITLE command in each syntax file (and you can't go wrong). All other commands should be placed below these commands.

Step 1.
I had suggested that you come up with 1 (or preferably 2) primary hypothesis(es) for your final paper (i.e. involving 1 or 2, independent variables). Run binary logistic regression with this independent variable, or two separate logistic regressions with each independent variable if you have two primary hypotheses.

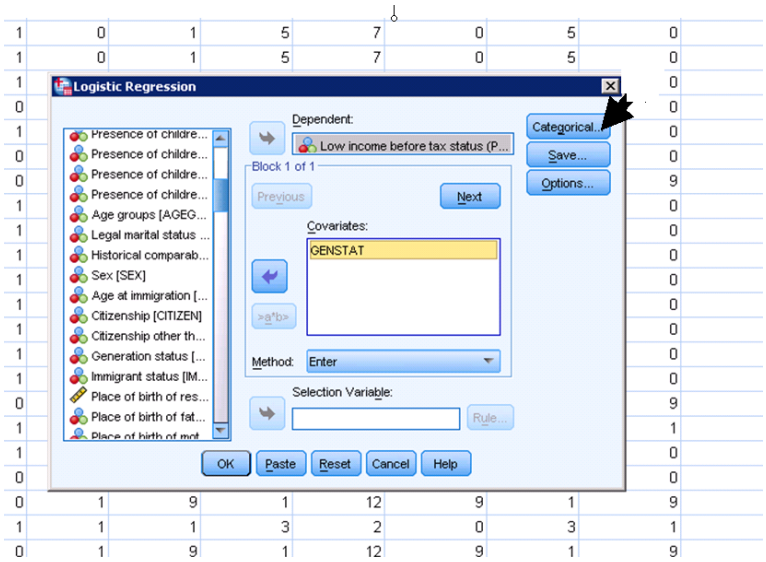
Note: This all relates to your literature review: which independent variable(s) in your proposed research do you consider most important or most interesting in explaining your dependent variable. What do you want to build your analysis around? What are you trying to highlight in your research paper?

As mentioned in class, when working with "Binary logistic regression" you can work with variables measured at various levels of measurement, for example, you can enter directly into logistic regression variables measured at the interval/ratio level, ordinal and nominal, in other words, it is not necessary to create "dummy variables" in Logistic Regression. But it is important to note that when you enter "nominal" or "crudely categorized ordinal variables, you MUST specify that they are to be treated as "categorical covariates" and MUST also identify a reference category (for ease of interpretation).

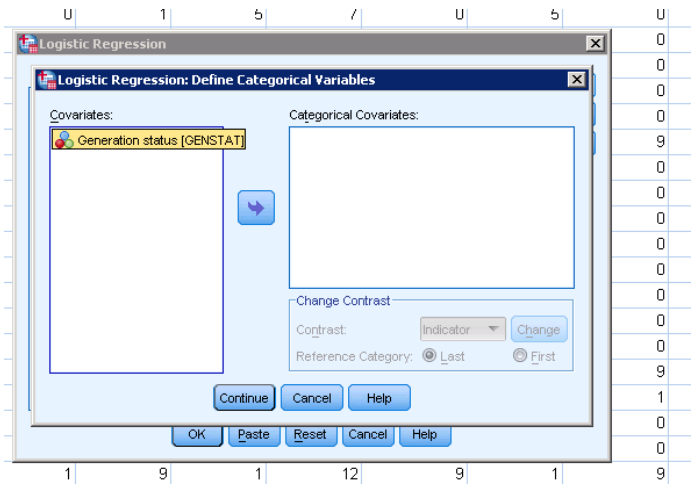
Let us take an example: assume that we are interested in examining the impact of Generational status (GENSTAT) on the "likelihood of low income" in Canada (our dependent variable is 0 – not low income; 1 low income). GENSTAT is coded as:

1. first generation, born in another country
2. second generation, both parents born abroad
3. second generation, one parent born abroad
4. third or higher generation

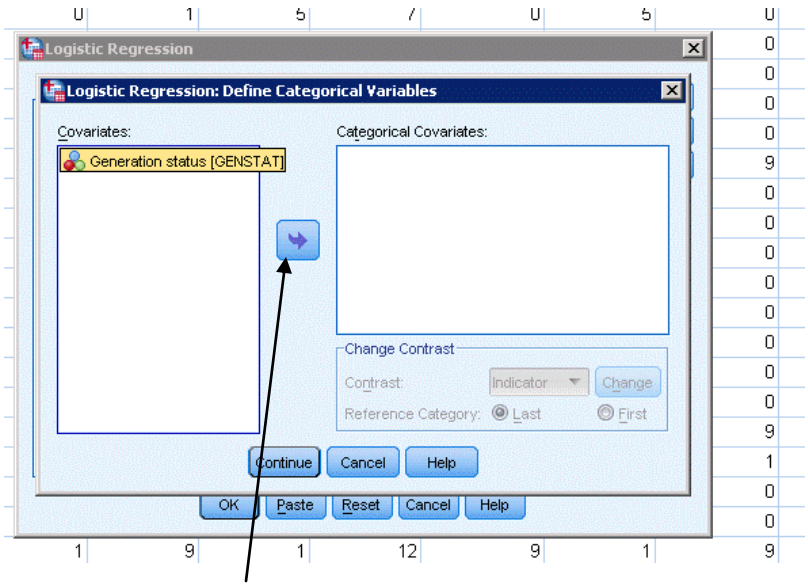
With logistic regression, we directly enter this variable GENSTAT into the binary logistic procedure (even though it is nominal). In the following example (below), we run Regression>Binary Regression, to obtain the following window where we identify the relevant dependent variable (low income) and our relevant independent variables (GENSTAT). Once we have done so, we must always click on the "Categorical" button when we work with nominal variables (we must explicitly tell SPSS how to handle this variable).



After clicking on the categorical button, we are given a window with two boxes:

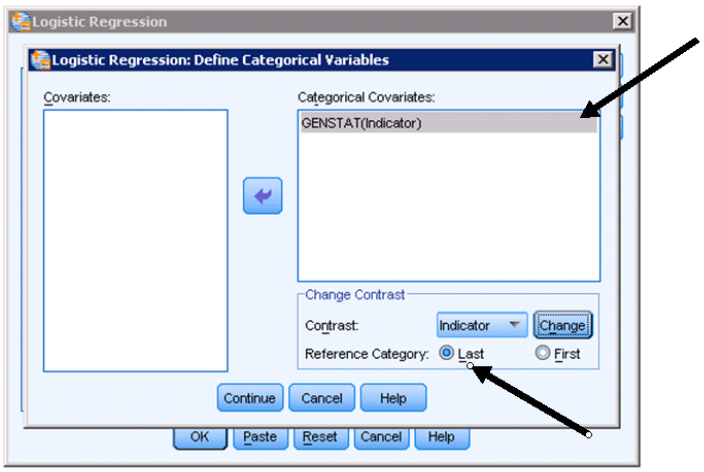


By default (unless specified otherwise) Logistic regression places all variables in the covariates box. In logistic regression, only interval ratio and ordinal variables can remain in that box.



We can use the arrow icon to move variables from one box to the other: In logistic regression, nominal and crudely categorized ordinal variables must always be placed in the "Categorical covariates" box, whereas other variables (ratio/interval variables and many ordinal variables) can remain in the "covariates box".

In the below box, we have moved "GENSTAT" appropriately into the "Categorical covariate" box (necessary since it is measured at the nominal level).



With all our "categorical covariates" we must specify a "reference category". This is not necessary with other variables in Logistic Regression (i.e. those treated as covariates). The two options available in specifying a "reference category" is to assign the first category as the reference category or the last category in the original varaible. In this example (see above), we select the last category of GENSTAT as our reference (according to the aforementioned coding

on this variable, that would be persons of the (4) category, i.e. persons who have had both parents born in Canada, i.e. 3rd generation or higher. The selection of "relevant reference category" should be done after inspecting the coding of the variables involved, but it is rather arbitrary. It is also fundamental in interpreting the effect of the variables involved.

The selection of a reference category is important in interpreting the results of logistic regression. After running the above regression (GENSTAT and low income), the subsequent odds ratios on this variable (GENSTAT) is obtained. It will only make sense in reference to this reference category.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GENSTAT			8135.148	3	.000	
	GENSTAT(1)	.623	.008	6865.845	1	.000	1.864
	GENSTAT(2)	-.125	.014	81.012	1	.000	.882
	GENSTAT(3)	-.134	.014	86.228	1	.000	.875
	Constant	-1.887	.005	169453.031	1	.000	.151

a. Variable(s) entered on step 1: GENSTAT.

In reading the above table, our reference category is 3rd generation or higher. GENSTAT(1) makes reference to category 1 with the old variable "first generation", GENSTAT(2) makes reference to the next category "second generation, both parents born abroad", GENSTAT(3) makes reference to the subsequent category "second generation, one parent born abroad". The corresponding "odds ratios" tell us how subsequent generations are doing, all relative to our reference category (3rd generation or higher).

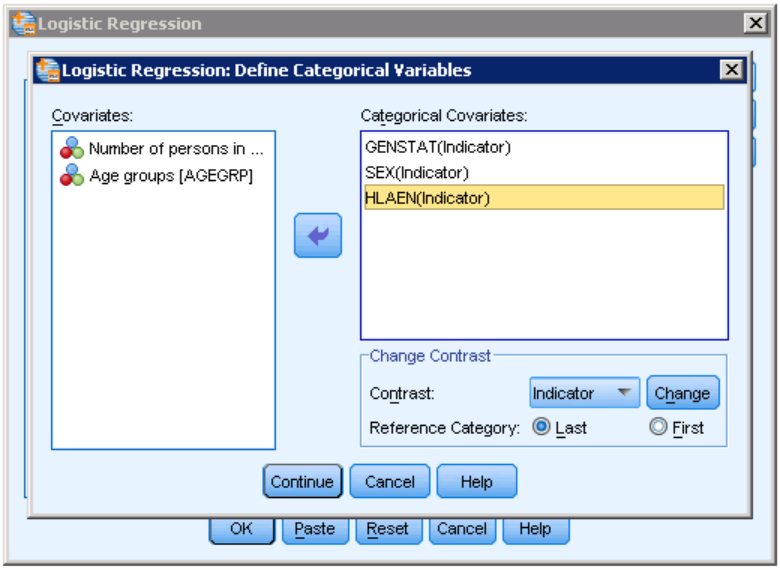
Quite clearly, the first generation are experiencing considerable hardship (the odds of falling into poverty are 86.4 per cent higher among the 1st generation than for the 3rd plus generation, i.e. (1.864-1.0)*100. The results for the second generation categories are quite encouraging, as both categories are "less likely" to fall into poverty than the 3rd generation. As merely an example, the GENSTAT(2) odds ration tells us that the "second generation, with both parents born abroad" have a lower odds of falling into poverty, in fact, their odds are 11.8 per cent lower than the 3rd generation, i.e. (0.875- 1.0)*100.

It is recommended to work with this categorical option whenever working with nominal variables in logistic regression and/or when introducing crudely categorized ordinal variables (i.e. for example, ordinal variables that have only three or four response categories). Do note that with other variables, you can merely introduce them into the model as "covariates" with no need to specify a reference categories (i.e. you need not specify reference categories for interval ratio and ordinal variables.

Highlight with a yellow marker on your output: (i) the odds ratio associated with each variable, (ii) the slope associated with each variable, (iii) whether or not your variable has a significant effect or not (at the .05 level), (iv) the odds of your reference category, and (v) the overall R^2 of each regression (in this case, we will work with the pseudo R^2 as specified in class, i.e. Nagelkerke R^2.

The only stuff to hand in with Step 2 are your syntax and output files.

Step 2. Move beyond simple regression, to include all independent variables simultaneously into a multiple regression. In your research proposal, I had suggested that you also consider potential control variables in your multivariate model (three of four). Note: your model may involve both "categorical covariates" and "covariates". Again, with all "Categorical covariates" you must specify the relevant reference category whereas it is not necessary with "covariates". For example, in this example we included 5 independent variables overall (two interval/ratio variables, age and # of persons in household as "covariates") and 3 variables as "categorical variables" (GENSTAT, SEX and Home language; HLAEN). With the latter three variables we would have to specify reference categories. See lecture notes for the further guidance on how to interpret the odds ratios in logistic regression when involving several variables simultaneously.



With the results from your regression, again highlight with a yellow marker on your output : (i) the odds ratio associated with each variable, (ii) the slope associated with this regression, (iii) whether or not your variable(s) have a significant effect or not (at the .05 level), (iv) the odds of your reference category, and (v) the overall R^2 of the regression (Nagelkerke).

If you anticipate the possibility of an interaction effect, it is here that you will have to introduce an interaction term into your model (as will be discussed in class). This usually involves computing a new variable which is the product of the two independent variables hypothesized as interacting. For example, assume that you hypothesize an interaction between VAR1 and VAR2 in explaining your dependent variable VAR3.

You first need compute an interaction term, as:

COMPUTE NEWVAR = VAR1 * VAR2.

You can then include all three variables in your regression (i.e. VAR1, VAR2 and NEWVAR). If NEWVAR has a significant effect, then this is supportive of the hypothesis of interaction. In including this interaction term, it doesn't matter if VAR1 and VAR2 continue to have a

significant effect, but it is important that you continue to include them in the model along with the interaction term (i.e. the final regression must include VAR1, VAR2 and NEWVAR).

Note: You need not "hypothesize" or "test" for interaction effects in this assignment or in your final paper, although it can be interesting. Recall the example provided in class, i.e. the interaction effect of "immigrant status" and "education" in explaining income, i.e. the pay-back for education tends to be much lower for immigrants than it is for other Canadians. Introducing an interaction term which is the product of "immigration status" and "education" could explicitly test for this.

Only stuff to hand in are your syntax and output files.

Step 3

After completing these regressions, I would like you to present the results from all of your regressions in steps 1 to 2 in a separate table. The table should include the regression results obtained from steps 1 and 2.

In so doing, I suggest that provide me with a Table that resembles the format as specified below (Table 2). The coefficients included are the "odds ratios" reported in SPSS.

Table 2. Low Income, Logistic Regression with Selected Independent Variables, 2001 Census

Covariates	Model 1	Model 2	Model 3
Generation (i)			
1st generation	1.855 ***		2.221 ***
2nd generation	0.873 ***		0.992 ***
Education		0.902 ***	0.922 ***
Age			0.975 *
Region (i)			
Atlantic Canada			1.21 ***
Quebec			1.01
Ontario			0.978 ***
Prairies			0.988 *
Household size			0.991
Sex (i)			1.284 ***
N	738333	738333	722444
Negalkerke's R^2	0.02	0.033	0.121

P-values ; * P-value < .05; ** P-value < .01 *** P-value < .001

(i) Reference category: 3rd generation or higher; British Columbia, and male

A nice software for creating these tables is "Excel" (it is very easy to use), or alternatively merely use the Table function available in "Word". The coefficients as presented in this Table are the "ODDS ratios" as obtained from SPSS.

Present the results from Steps 1 and 2 as separate models. In the first model, present the results from the logistic regression for one of your independent variables. In the second model, include the results from the second regression with the other independent variable (if applicable). Regardless of whether you work with one or two regressions, you must also present an additional model (model 3 in our example) with all the variables in your model, including all relevant control variables (this is your full model). The simpler models are called "nested" models, relative to this complete model with all relevant explanatory variables.

This table should also specify the sample sizes of all of your regressions. To locate the sample size of your regression in your SPSS output, it is possible to locate the number in the Case Processing Summary that accompanies each logistic regression. The N is equivalent to the "cases involved in the analysis". This is not necessarily the same number of cases across the different regressions, due to the issue of missing cases (see note below).

As output from Part C all I am asking for is your syntax files and corresponding output files, organized as steps 1 through 3. I am also asking for this summary Table without any interpretation. The results from the "weighted regressions" in Part C can be used in your final paper.

In your final paper, you will have to make sense of this table. What do they tell us about the initial relationships as hypothesized? Are they significant? Do you find support for your initial hypotheses? Do they remain, even after including all your relevant control variables??

Note: if you wish, you can depart a bit from the above recommended steps, by including more than 3 models (perhaps 4 alternatives, or even 5 nested models). You might for example, in addition to what is recommended above, include an additional model with the two independent variables together, or two variables together with an interaction term, all prior to presenting all of the variables together in your full model. I will allow for some flexibility on this, if you so chose.

NOTE: For the current assignment, there is no need for interpretation (merely put together the two tables). THAT'S IT!!! Interpret these tables in your final paper.

A few additional observations:

In our empirical research this year we have yet to explicitly address the problem of "missing values". In a database, "missing values" can surface as a direct result of a variety of factors, including the possibilities of "don't know", "refusal" or "non-applicable" in gathering information through a survey. Typically, non-applicable is not that great an issue as we can avoid this type of missing value by properly defining our subsample for analysis. On the other hand, "refusals" and/or "don't know" can potentially be problems.

The default procedure in regression when working with SPSS is to merely delete "missing cases" on a "LISTWISE" basis. Listwise deletion uses only cases from your data set that have valid (non-missing) values for all variables included in the regression. While this is the default procedure in regression, it is not always the preferred option. The problem with "Listwise"

deletion of missing cases is that you can potentially lose quite a few cases (and this can seriously reduce your sample size). If you find that you are losing a lot of cases, you should be clear on why this is occurring. Can it be avoided? If not, do note it when reporting your final results in your final paper.

NOTE: If you are working with the CENSUS, missing values is not an issue (i.e. there are not any in this dataset, as Stats Canada imputes all missing information). The three regressions should be identical when working with the Census, unless you have made some sort of serious mistake in terms of "not applicable".

Also, for future reference:

It is useful to examine initially your data with unweighted counts. This provides you with some sense as to potential problems with small numbers and missing cases.

Yet if you require descriptive estimates on how the variables likely look in the population (example, frequency distributions), then run all your procedures with the correct sample weights. If your sample design is not simple random sample, then these results can be quite different from what was initially observed. On the other hand, if your survey is based on a "simple random sample", whereby all persons in the sample had an equal probability of selection, then your results from the "unweighted" procedure will likely be near identical to what you obtain with the above procedure (note: the 2006 public use file is a "simple random sample" of the complete census.

One final observation:

When you move into your final multivariate analysis, the only way to get meaningful significance tests is to revise the weight as specified above (this is not perfect, but it is certainly better than doing nothing). In using this method, the subsequent estimates should not be seriously biased and the significance tests should be reasonably accurate. This is a preferred option over not "weighting" your analysis or merely doing the analysis with the initial weights. There are more complex procedures possible, beyond the scope of the current course, that are preferred to the recommended procedure as specified above (save it for graduate school).