- **Assignment 3 is now posted (due: April 2nd)**

- **Complete ONE of the two assignments posted;**
- **either "OLS Linear" regression or "Logistic" regression.**

- **NOTE:**
- **RESULTS FROM ASSIGNMENT 3**
- **-> RESULTS SECTION FOR YOUR FINAL PAPER**

- **Final exam: Tuesday April 17th, 2:00 p.m. (LH103)**

- **Final paper (due in my office, a week after the last class, Monday April 16th, 5:00 p.m.)**

Today:

A few observations on Assignment 2

A few additional comments on “Binary Logistic Regression”
-> how to handle independent variables
(categorical covariates; covariates)
-> Nagelkerke R^2
-> Hosmer-Lemeshow Goodness of Fit index

Next class:
Tips on creating models, regardless of whether we are working with OLS or Logistic Regression..

- A few observations on Assignment 2

- Dependent variable:
  # of outings (per month)..

- Interval ratio dependent variable..
- perfect for OLS regression..

- 3 independent variables..
- Sex;
- # of close friends/relatives;
- Marital Status

- With OLS regression, MUST create “dummy variables” with “nominal variables”..

```
RECODE SEX (1=1) (2=0)  INTO REC_SEX.
VARIABLE LABELS  REC_SEX 'Recoded Sex'.
EXECUTE.

RECODE MARSTAT (1=1) (2 thru 6=0) (8 thru 9=SYSMIS) INTO Married.
VARIABLE LABELS  Married 'married persons'.
EXECUTE.

RECODE MARSTAT (1=0) (2=1) (3 thru 6=0) (8 thru 9=SYSMIS) INTO COMLAW.
VARIABLE LABELS  COMLAW 'common law '.
EXECUTE.

RECODE MARSTAT (3=1) (1 thru 2=0) (4 thru 6=0) (8 thru 9=SYSMIS) INTO Widow.
VARIABLE LABELS  Widow 'widowed person'.
EXECUTE.

RECODE MARSTAT (6=0) (1 thru 3=0) (4 thru 5=1) (8 thru 9=SYSMIS) INTO SEPDIV.
VARIABLE LABELS  SEPDIV 'separated/divorced'.
EXECUTE.

RECODE MARSTAT (6=1) (1 thru 5=0) (8 thru 9=SYSMIS) INTO SINGLE.
VARIABLE LABELS  SINGLE 'single persons'.
EXECUTE.

REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT NUMEVACT  /METHOD=ENTER Married COMLAW Widow SEPDIV REC_SEX ISL_Q020
```

Sex
0. female;   1 male

Married
0. no;   1 yes

Common law
0. no;   1 yes

Widowed
0. no;   1 yes

Sep/divorced
0. no;   1 yes

Single
0. no;   1 yes

All dummies, except for “SINGLE”;

Also sex, number of close relatives

NOTE THIS IS NOT IDENTICAL TO YOUR ASSIGNMENT (DIFFERENT SUB-SAMPLE)

$R^2 = 0.106$ .. Pretty good, right?
IV's Explain over 10 percent of the Variance in our dependent variable..

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.326[a]	.106	.106	15.615

a. Predictors: (Constant), Number of close relatives and friends who live in the same city or community, Recoded Sex, common law , separated/divorced, widowed person, married persons

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	526519.804	6	87753.301	359.910	.000[b]
	Residual	4429967.509	18169	243.820		
	Total	4956487.312	18175			

a. Dependent Variable: Average number of evening activities respondent goes out for in a month

b. Predictors: (Constant), Number of close relatives and friends who live in the same city or community, Recoded Sex, common law , separated/divorced, widowed person, married persons

all significant;
p-value < .001

Marital status seems relevant

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	27.106	.281		96.494	.000
	married persons	-9.447	.299	-.286	-31.629	.000
	common law	-5.157	.441	-.094	-11.696	.000
	widowed person	-17.574	.472	-.298	-37.245	.000
	separated/divorced	-8.246	.423	-.159	-19.476	.000
	Recoded Sex	3.098	.237	.093	13.094	.000
	Number of close relatives and friends who live in the same city or community	.108	.008	.091	12.938	.000

a. Dependent Variable: Average number of evening activities respondent goes out for in a month

Excluded
Single
(reference)

Sex
0. female
1. male

Married persons go out 9.4 times fewer than Singles

Men are going out more so then women.. 3.098 times more..

Many friends/relatives encourage outings..
with each addition person in network, predict .108 additional outings

- What have we concluded?

Richmond Street, London ON

- What have we concluded?

- Enjoy yourself while you still can…

- Working with Binary Logistic Regression

- Dependent variable:
- Smoking behavior..
- 0 - no
- 1 - daily smoker

```
DATASET ACTIVATE DataSet1.
USE ALL.
COMPUTE filter_$=(DHHGAGE GE 7).
VARIABLE LABELS filter_$ 'DHHGAGE GE 7 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

For the purpose of this assignment we created 4 dichotomous variables

```
RECODE SMK_202 (1=1) (2=0) (3=0) (7=SYSMIS) (8=SYSMIS) (9=SYSMIS) INTO
SMOKE.
VARIABLE LABELS SMOKE 'Smoker'.
EXECUTE.

RECODE EDUDR04 (1=0) (2=0) (3=1) (4=1) (7=SYSMIS) (8=SYSMIS) (9=SYSMIS) INTO
EDUCATION.
VARIABLE LABELS EDUCATION  "'postsecond grad'.
EXECUTE.
```

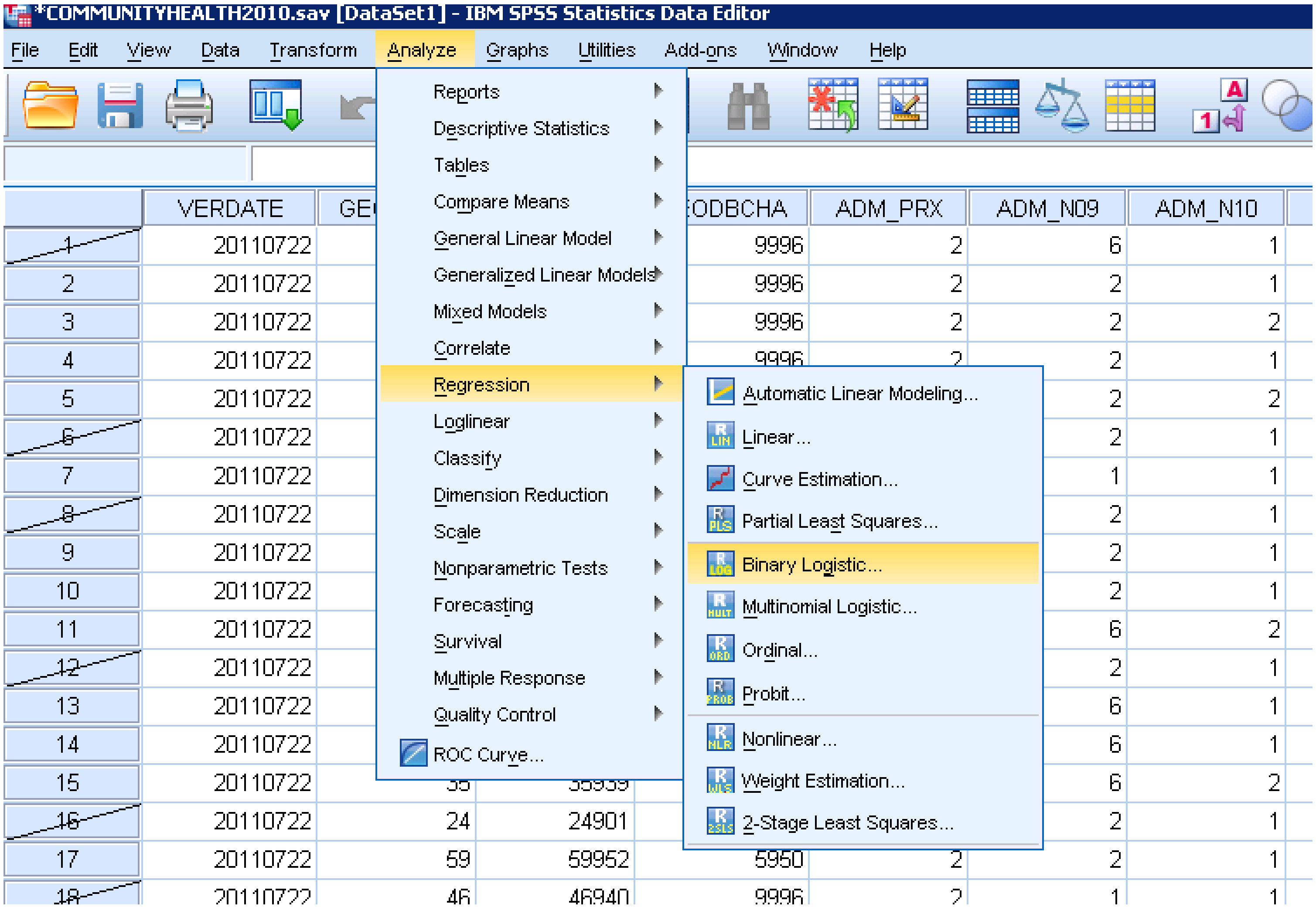
Post secondary grad
0 – no; 1 - yes

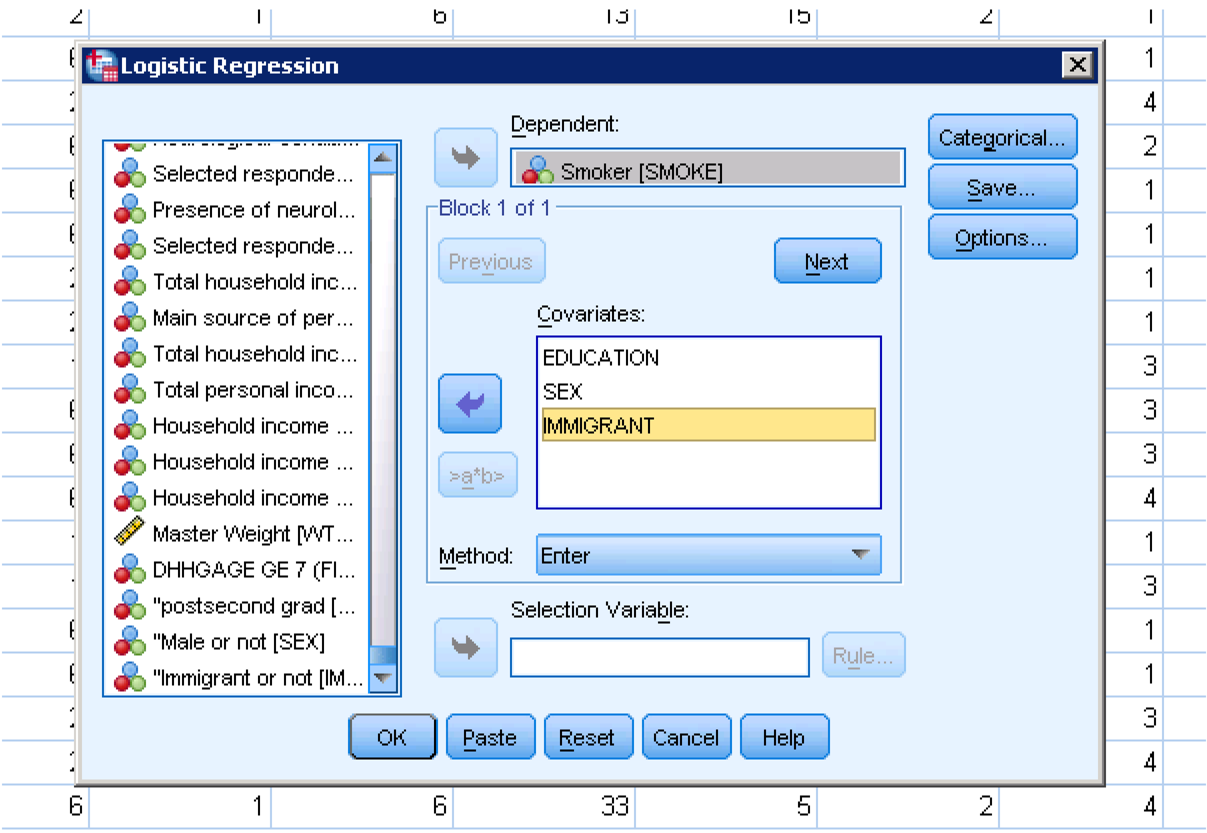
```
RECODE DHH_SEX (1=1) (2=0) INTO
SEX.
VARIABLE LABELS SEX "'Male or not'.
EXECUTE.
```

Sex
0- female; 1- male

```
RECODE SDCFIMM (1=1) (2=0) (7=sysmis) (8=sysmis) (9=sysmis) INTO
IMMIGRANT.
VARIABLE LABELS IMMIGRANT "'Immigrant or not'.
EXECUTE.
```

Immigrant
0- no; 1- yes


*COMMUNITYHEALTH2010.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
VERDATE
ADM_PRX
ADM_N09
ADM_N10
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...



Logistic Regression
Selected responde...
Presence of neurol...
Selected responde...
Total household inc...
Main source of per...
Total household inc...
Total personal inco...
Household income ...
Household income ...
Household income ...
Master Weight [WT...
DHHGAGE GE 7 (FI...
"postsecond grad [...
"Male or not [SEX]
"Immigrant or not [IM...
Dependent:
Smoker [SMOKE]
Block 1 of 1
Previous
Next
Covariates:
EDUCATION
SEX
IMMIGRANT
>a*b>
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
OK
Paste
Reset
Cancel
Help

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	38157.967[a]	.014	.023

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Sex
0- female; 1- male

Education
0 - not a grad; 1 - post sec grad

Immigrant
0 – no; 1 - yes

Classification Table[a]

Observed			Predicted Smoker .00	Predicted Smoker 1.00	Percentage Correct
Step 1	Smoker	.00	35117	0	100.0
		1.00	7242	0	.0
	Overall Percentage				82.9

a. The cut value is .500

All variables have a significant effect.. P < .001

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SEX	.313	.026	145.266	1	.000	1.368
	EDUCATION	-.312	.026	140.811	1	.000	.732
	IMMIGRANT	-.681	.043	255.408	1	.000	.506
	Constant	-1.452	.024	3805.837	1	.000	.234

a. Variable(s) entered on step 1: SEX, EDUCATION, IMMIGRANT.

Sex; men have 36.8 percent higher odds of smoking (1.368-1.0)*100 relative to women..

Education; post-secondary grads have 26.8 percent lower odds of smoking relative to non-grads, i.e. (0.732 - 1.0) *100 = -26.8%

Immigration status; immigrants have 49.4 percent lower odds of smoking relative to non-immigrants, i.e. (0.506 - 1.0) * 100 = 49.4%

- Demographers speak of the:
- “Healthy Immigrant effect”..
- Populations with higher percentage immigrant in Canada tend to be healthier..

- Our results tend to suggest that the healthiest would be “immigrants” who are female and well educated…

http://www.statcan.gc.ca/pub/11-633-x/11-633-x2016003-eng.pdf

Catalogue no. 11-633-X — No. 003
ISSN 2371-3429
ISBN 978-0-660-06457-4

Analytical Studies: Methods and References

The 2001 Canadian Census–Tax–Mortality Cohort: A 10-Year Follow-up

by Lauren Pinault, Philippe Finès, Félix Labrecque-Synnott, Abdelnasser Saidi, and Michael Tjepkema

Release date: October 26, 2016

Statistics Canada Statistique Canada

Canada

Table 2
Remaining life expectancy at age 25, by sex and selected socioeconomic and demographic variables

Category	Total			Men			Women		
		95% confidence interval			95% confidence interval			95% confidence interval	
	Years	From	To	Years	From	To	Years	From	To
	number								
Total	**56.8**	**56.8**	**56.9**	**54.6**	**54.5**	**54.6**	**59.0**	**59.0**	**59.1**
Educational attainment									
University degree	59.8	59.7	59.9	58.6	58.5	58.8	61.8	61.6	62.1
Postsecondary non-university certificate or diploma	59.3	59.2	59.4	56.7	56.5	56.8	60.8	60.7	61.0
High school with or without trades certificate	57.1	57.0	57.1	54.8	54.7	54.9	59.5	59.4	59.6
Less than secondary school graduation	54.4	54.3	54.4	51.9	51.8	52.0	56.8	56.7	56.9
Difference = university minus less than secondary school	5.4	5.4	5.5	6.7	6.7	6.8	5.0	4.9	5.1
Income adequacy quintile (area)									
5 (highest)	58.9	58.8	59.0	57.4	57.2	57.5	60.8	60.6	60.9
4	57.9	57.8	58.0	55.9	55.8	56.0	60.2	60.0	60.3
3	57.1	57.0	57.2	55.0	54.9	55.1	59.5	59.4	59.6
2	56.0	55.9	56.1	53.4	53.3	53.5	58.6	58.4	58.7
1 (lowest)	53.8	53.7	53.8	50.5	50.4	50.6	56.2	56.1	56.3
Difference = quintile 5 minus quintile 1	5.2	5.2	5.2	6.8	6.8	6.8	4.5	4.5	4.6
Aboriginal identity									
No Aboriginal identity	57.1	57.1	57.2	54.9	54.9	55.0	59.3	59.2	59.3
Any Aboriginal identity	50.1	49.9	50.3	49.0	48.7	49.3	52.3	52.0	52.6
North American Indian identity only	49.7	49.4	49.9	47.5	47.2	47.9	51.8	51.5	52.2
Métis identity only	52.9	52.4	53.4	50.7	50.1	51.3	55.2	54.5	55.9
Inuit identity only	46.5	45.9	47.2	45.2	44.3	46.1	47.8	46.9	48.8
Difference = not Aboriginal minus Aboriginal	7.0	7.2	6.9	5.9	5.7	6.1	7.0	6.7	7.3
Visible minority status									
Not a visible minority	56.8	56.8	56.9	54.6	54.5	54.6	59.0	59.0	59.1
Visible minority	60.8	60.6	60.9	58.9	58.7	59.1	62.5	62.3	62.7
Chinese	61.9	61.6	62.1	59.9	59.6	60.3	63.6	63.2	64.0
South Asian	60.0	59.7	60.4	58.9	58.3	59.4	61.4	60.9	62.0
Black	59.6	59.2	60.1	57.2	56.7	57.8	61.3	60.7	61.9
Filipino	60.1	59.6	60.6	57.4	56.7	58.2	61.9	61.2	62.6
Latin American	60.4	59.5	61.4	57.1	56.1	58.1	62.6	61.3	63.9
Southeast Asian	61.8	60.5	63.1	59.4	58.3	60.5	63.3	61.3	65.3
Arab	59.5	58.6	60.4	57.7	56.7	58.8	62.9	61.1	64.7
Difference = visible minority minus not visible minority	3.9	3.8	4.0	4.3	4.2	4.5	3.5	3.3	3.6

Sources: Statistics Canada, 2001 Canadian census–tax–mortality cohort, derived from the 2001 Census of Population and the 2014 Amalgamated Mortality Database.

- MORE ON LOGISTIC REGRESSION:
- An issue with “logistic regression”…
- Recall that we must “dichotomize” our dependent variable in Logistic regression..

- What of the independent variables?
- In assignment 2, we worked with dichotomous independent variables for ease of introducing this method.. (smoking yes/no; Immigrant yes/no; Sex male/female;
- PS education yes/no)

- Yet in LOGISTIC regression
- How do we handle “independent variables that are not dichotomous”
- for example, “ethnicity” (with 7 categories) or “region” (with 12? categories)

Recall also:
In Linear regression: we have to work with “Dummy Variables” when we have independent variables that are either “nominal variables” or crudely categorized “ordinal variables”.

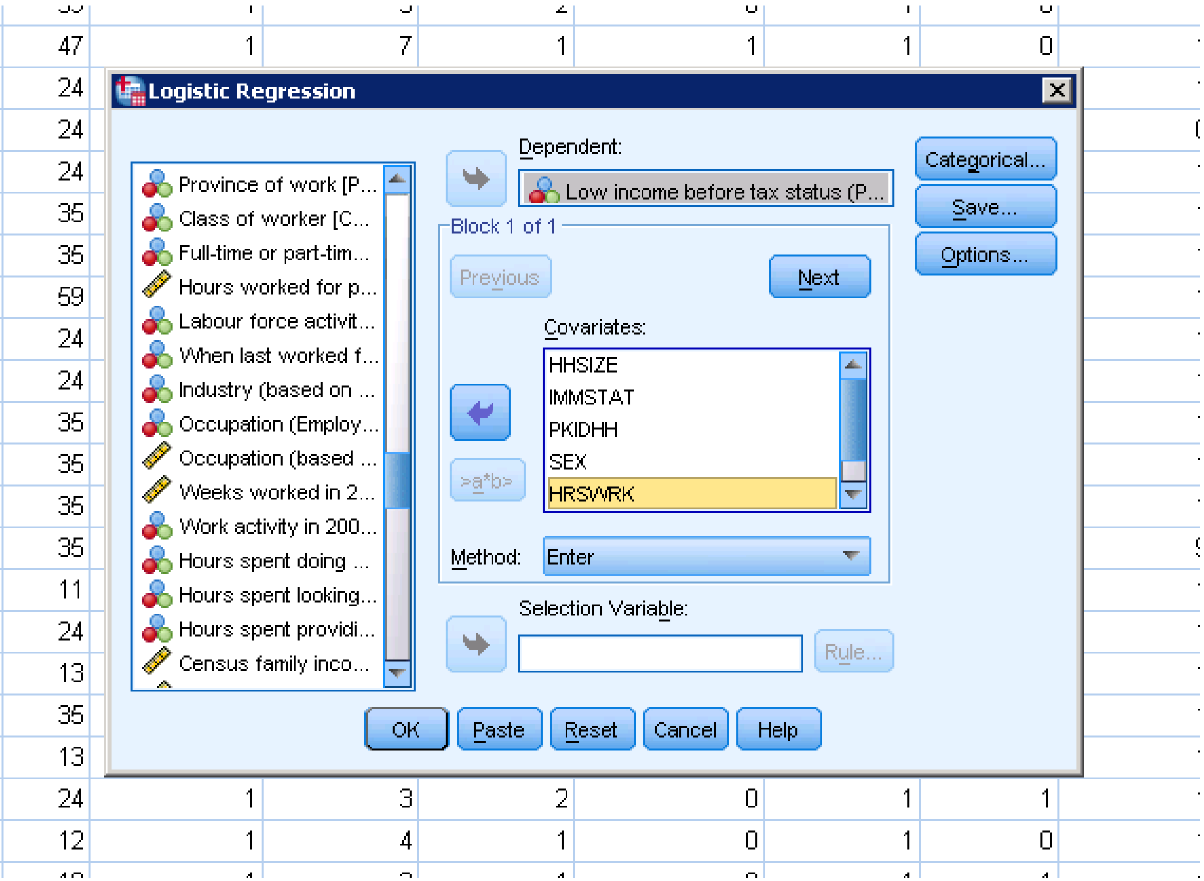
In Logistic regression:
WE DO NOT HAVE TO COMPUTE “DUMMY VARIABLES!

Yet in working with SPSS, we must carefully consider “level of measurement” of all of our independent variables and potentially specify “reference categories” for our analysis…

How so?

Let's select several independent variables, in the explanation of "low income"

Household size
Immigration status
Sex
Presence of children
Hours worked



Must think of level of measurement when running a logistic model

- In LOGISTIC regression, all types of variables can be directly used in the SPSS procedure:

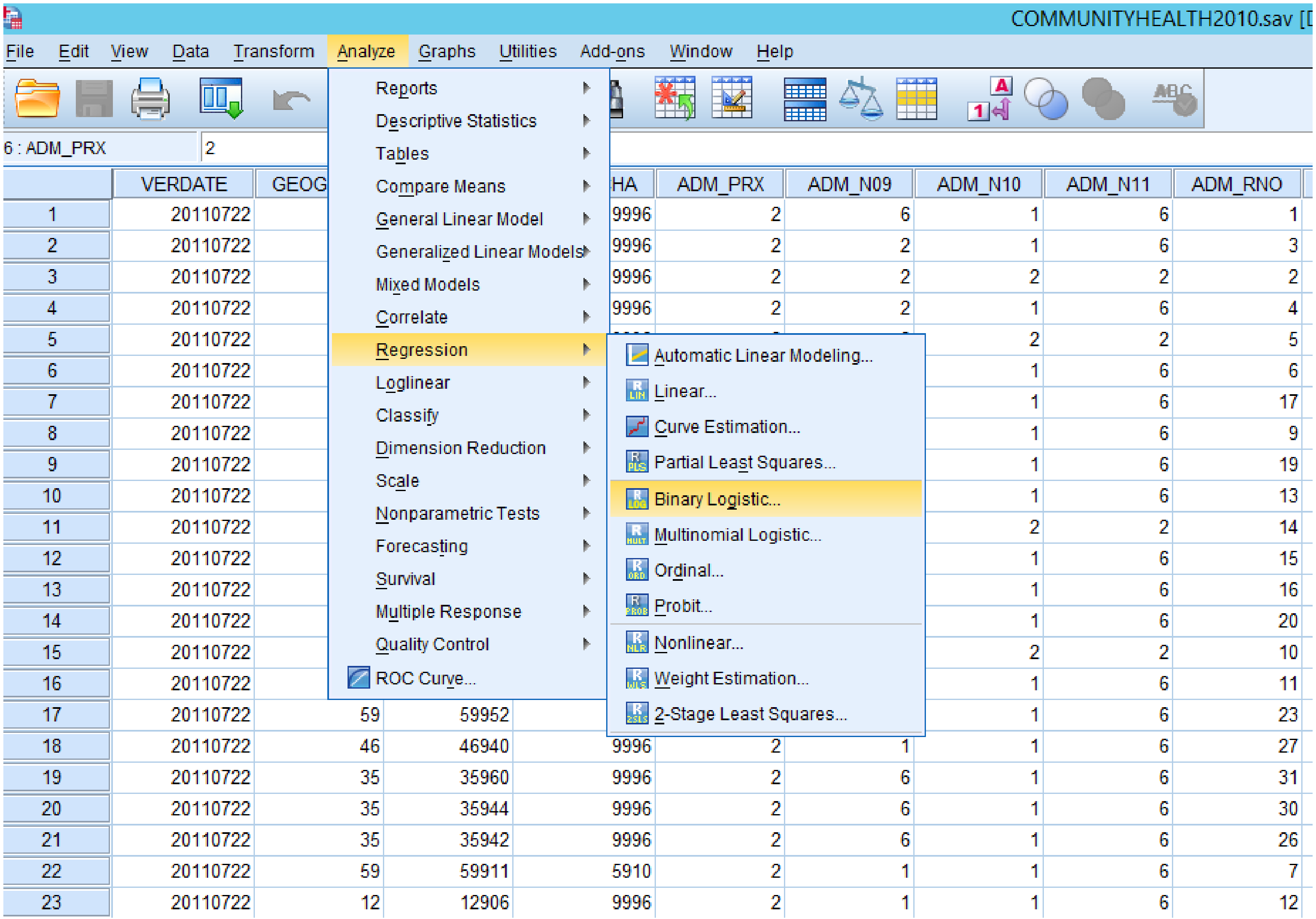
- it is merely necessary to identify “variables” as either a **“covariate”** or **“categorical covariate”**…

- **In logistic regression, we refer to:**

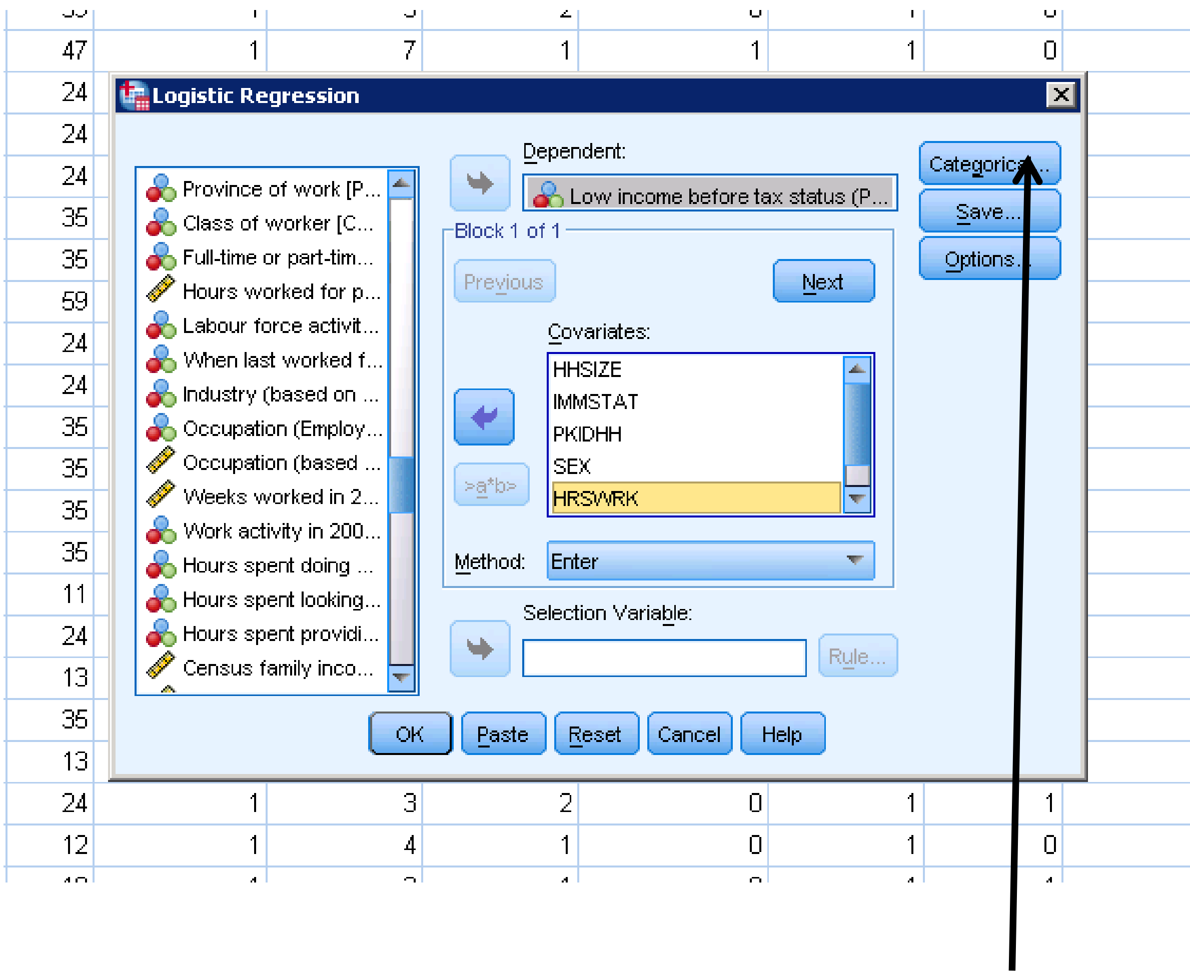
- **Covariates**: interval/ratio; ordinal variables

- **Categorical covariates**:
nominal variables” or crudely categorized “ordinal variables”.. (e.g. less than 5 categories)

Ex. Running a logistic regression on “low income” (0-no; 1-yes)



Let's select several independent variables, in the explanation of "low income"



This is where we assign variables as either "categorical covariates" or as "covariates"

INDEPENDENT VARIABLES		DEPENDENT VARIABLE
Household size	covariate	
Immigration status	categorical covariate	
Sex	categorical covariate	Low Income
Presence of children	categorical covariate	
Hours worked	covariate	
Province	categorical covariate	

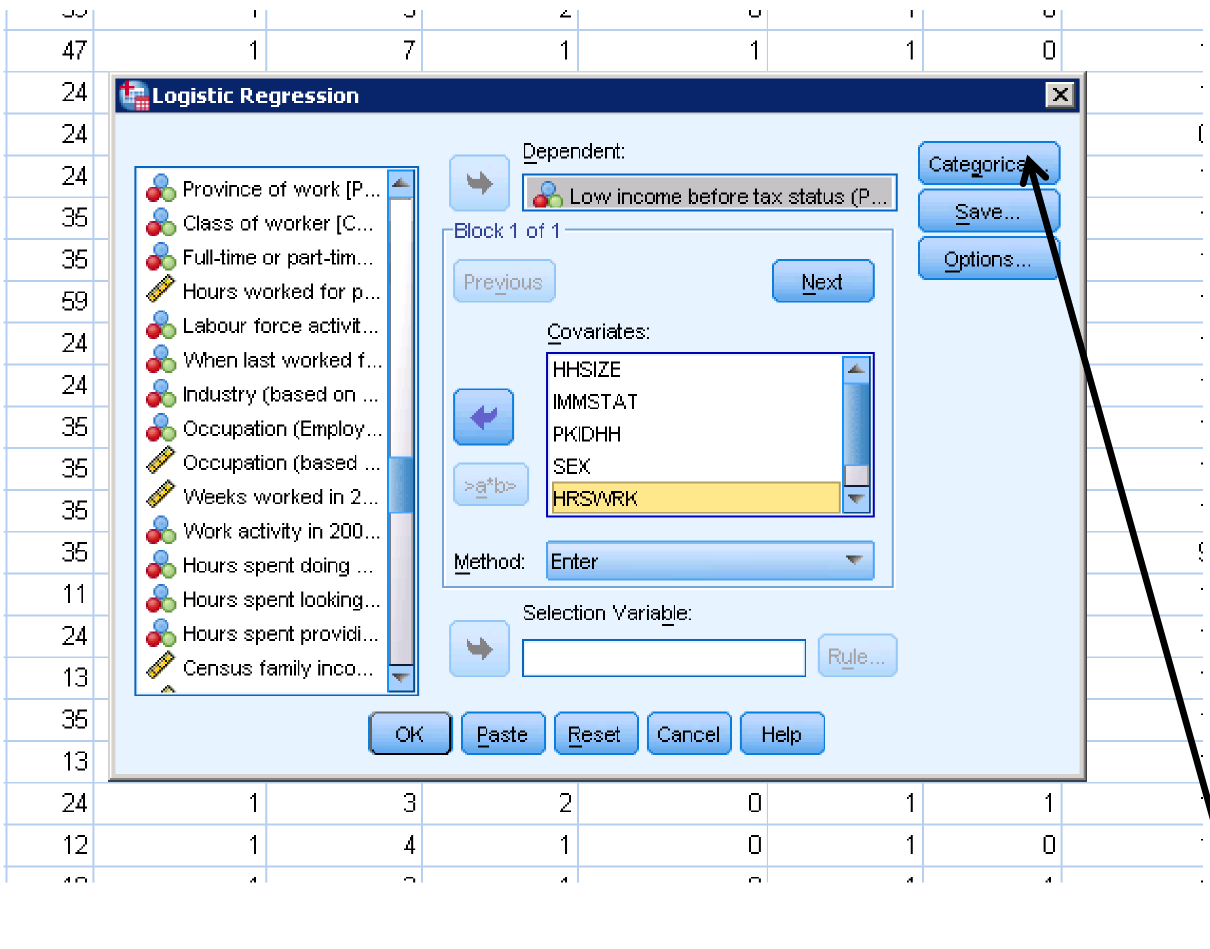
Covariate - interval/ratio; ordinal variables

Categorical covariate –
nominal variables" or crudely categorized "ordinal variables", with more than 2 categories

Note: if a nominal or ordinal variable is dichotomous (yes no; high low), you can actually treat it as a covariate or a categorial covariate.

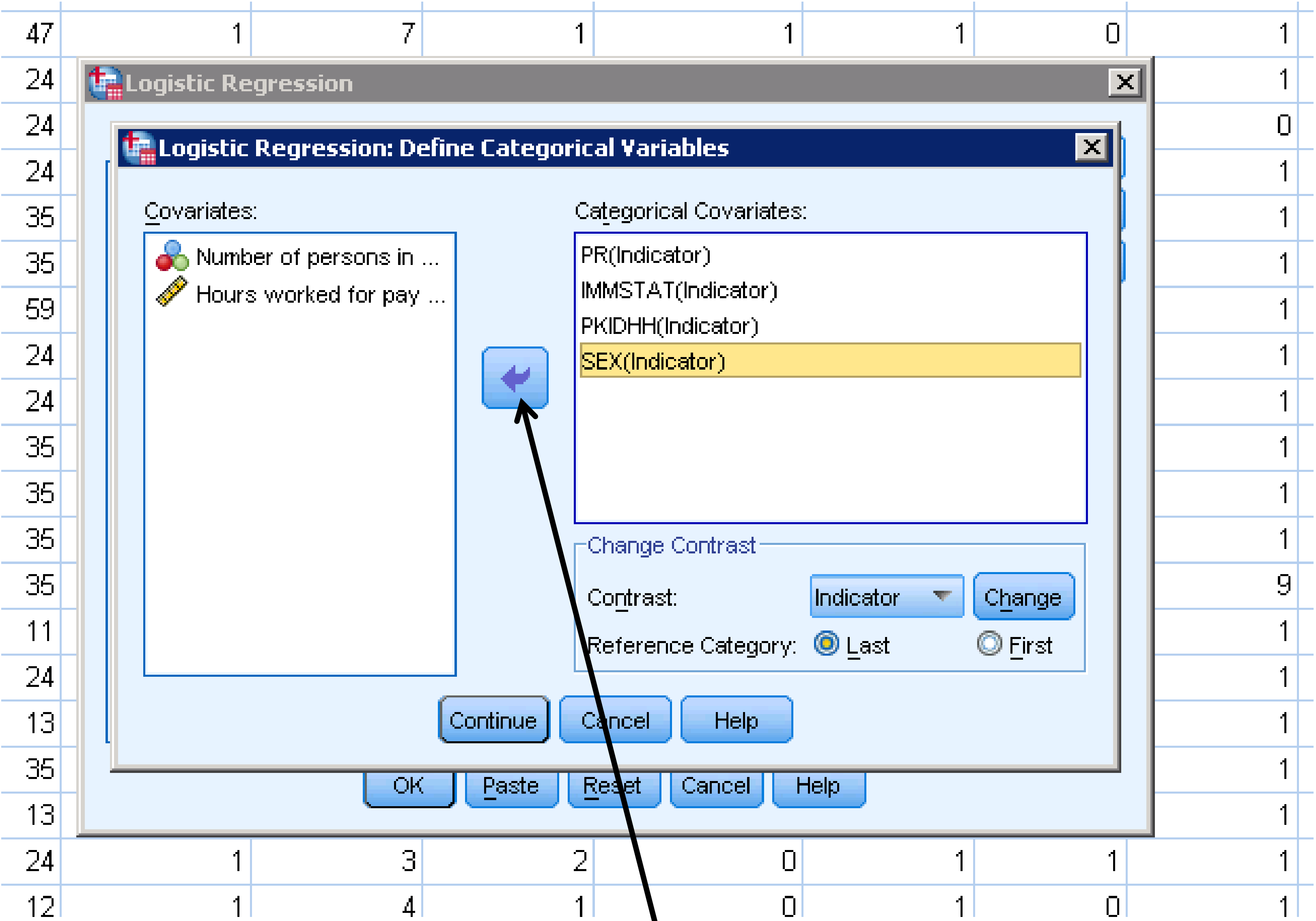
My rule of thumb: I only treat interval/ratio and ordinal variables as covariates.. Everything else, as a categorical covariate

Let's select several independent variables, in the explanation of "low income"



Assign variables as either "categorical covariates" or as "covariates"

Two boxes: covariates & categorical covaraites

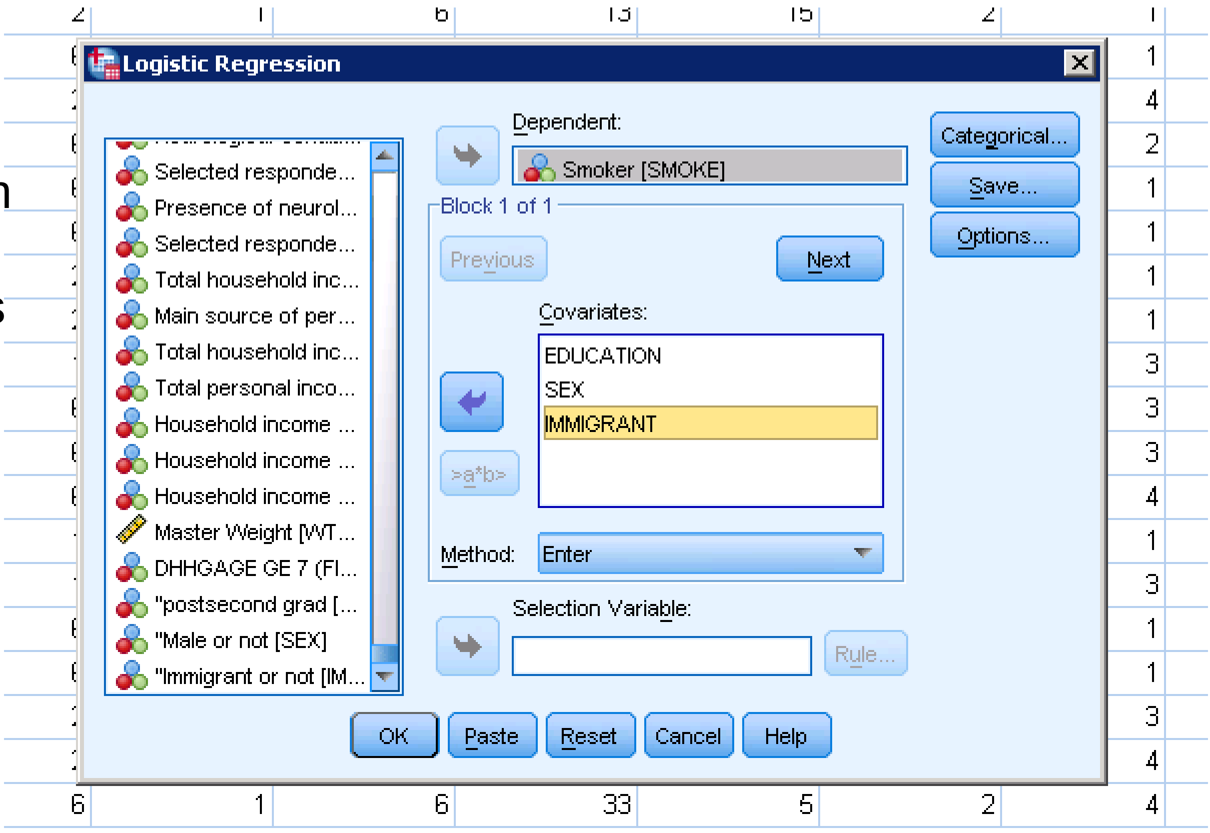


Default is "covariate"..

Can move back and forth across 2 boxes

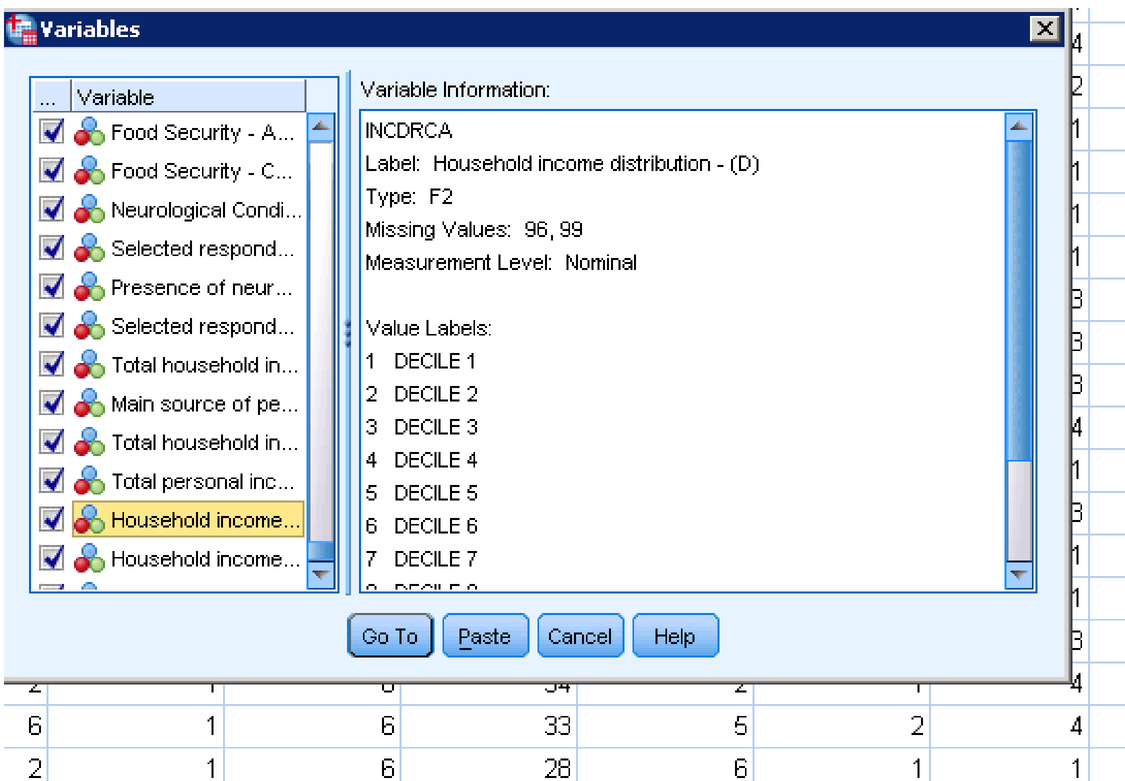
- Returning to the example from our assignment 2 .. on smoking behavior..
- How do we interpret "covariates" in
- Logistic regression??

We worked with dichotomous variables in this context

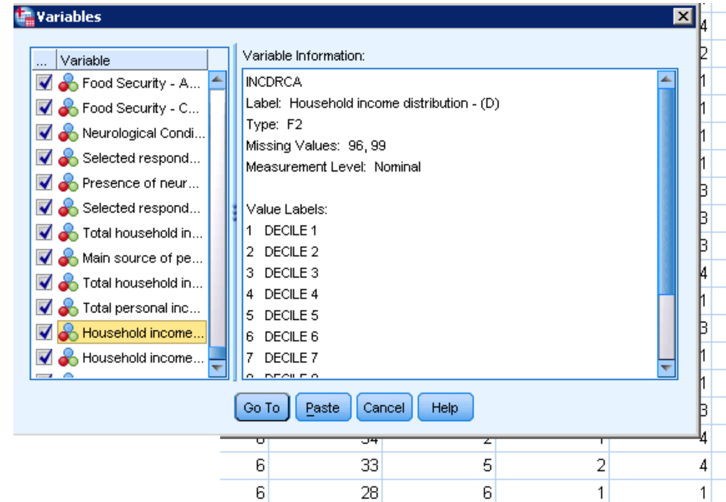


What if we added an additional variable:
Income decile of the respondent?
Does the respondent fall in the bottom 10 % of income earners, the second 10%,.. the top 10 per cent, etc?

This is an interval/ratio variable..
Must introduce it as a “covariate” and not a “categorical covariate”…



How to interpret?



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	EDUCATION	-.161	.030	29.013	1	.000	.851
	SEX	.336	.029	138.685	1	.000	1.399
	IMMIGRANT	-.722	.047	238.365	1	.000	.486
	INCDRCA	-.102	.005	373.135	1	.000	.903
	Constant	-1.001	.032	951.431	1	.000	.368

a. Variable(s) entered on step 1: EDUCATION, SEX, IMMIGRANT, INCDRCA.

Odds ratio

Income deciles variable

Significant P < .001

For each unit increase on our independent variable,
we expect the lower odds of smoking…

In moving into the next higher “income decile”, we would expect that the odds of smoking would be lower by 9.7 per cent (0.903 – 1.0) * 100

Returning to our Maple Leafs example:

Sex (0 – male; 1 – female)

Age (in years)

Toronto Resident
(0 – no; 1 – yes)

University Educated
(0 – no; 1 – yes)

Fan
0 – no
1 - yes

Obviously, more complex models are possible with many independent variables..

Sex: 0 male, 1 female
Age: years
Toronto Resident
0 no, 1 yes
Univ educated
0 no, 1 yes

$$\ln[p/(1-p)] = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

Dependent variable: Toronto Maple Leaf Fan (0 no, 1 yes)

e^b

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a] Sex	X1	b1 -.780	.124	39.624	1	.000	.458
Age	X2	b2 .020	.004	32.650	1	.000	1.020
Toronto Resident	X3	b3 1.618	.197	67.534	1	.000	5.044
University educated	X4	b4 -.023	.020	1.370	1	.242	.977
Constant		a -2.246	.363	38.224	1	.000	.106

a. Variable(s) entered on step 1: Sex Age Toronto Resident University educated

Which b's are significant?

Age is the only covariate: others are categorical, right?

? For each additional year of age, we expect the odds of being a fan to go up by about 2 per cent…(1.020 – 1.0)* 100

We must be careful in working with “categorical variables”..

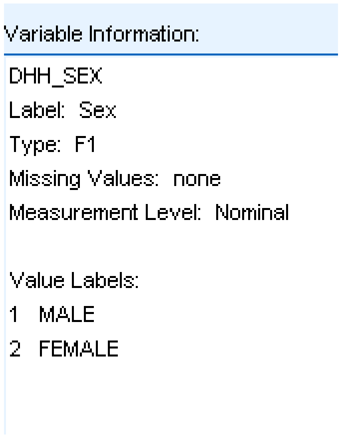
Nominal variables…

Last week, merely entered “dummy variables” as independent… and they were
Treated like any other variable (default, treated like a covariate).

There is a more preferred procedure…
Treat them as a “categorical covariate”, and specify reference category..

- Returning to our original smoking example,
- Considering exclusively Sex and Smoking

- Original independent variable



- How to work with "nominal variables" in Logistic Regression.

- With dichotomous variables
-

  Choice:
- You can either create and work with dummy variables, or
- You can enter your original variable directly without creating dummies **(recommended)**

- If the latter:

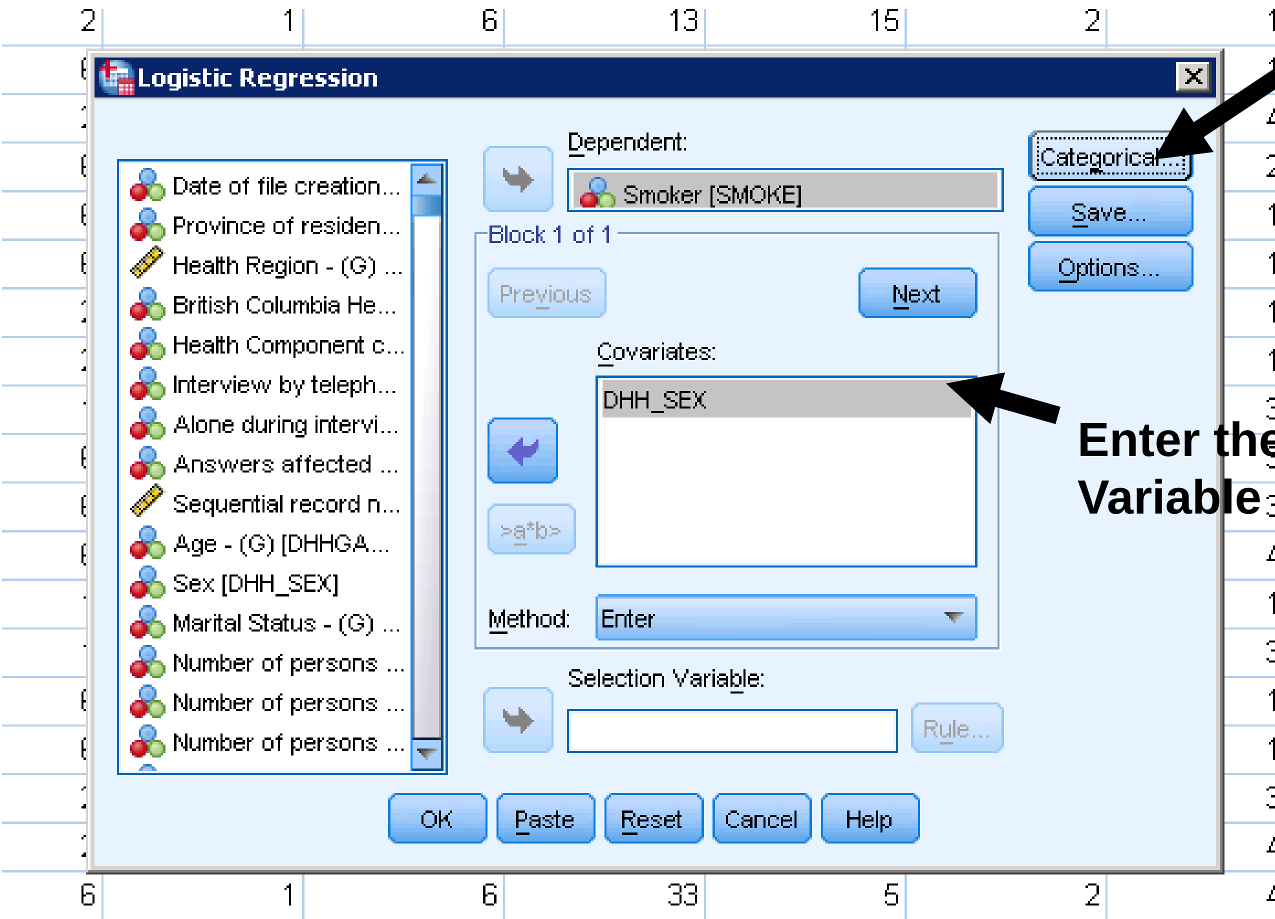
- 1. must always assign nominal variables as "**categorical covariate**" &

- 2. must identify a **reference category** for your analysis (details forthcoming)

- Example:

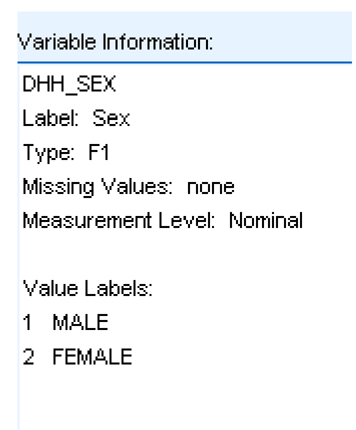
- Let's "not create" a dummy variable for sex,
- but merely enter the original variable into the logistic regression procedure"..

- Can merely introduce DHH_SEX into our logistic model

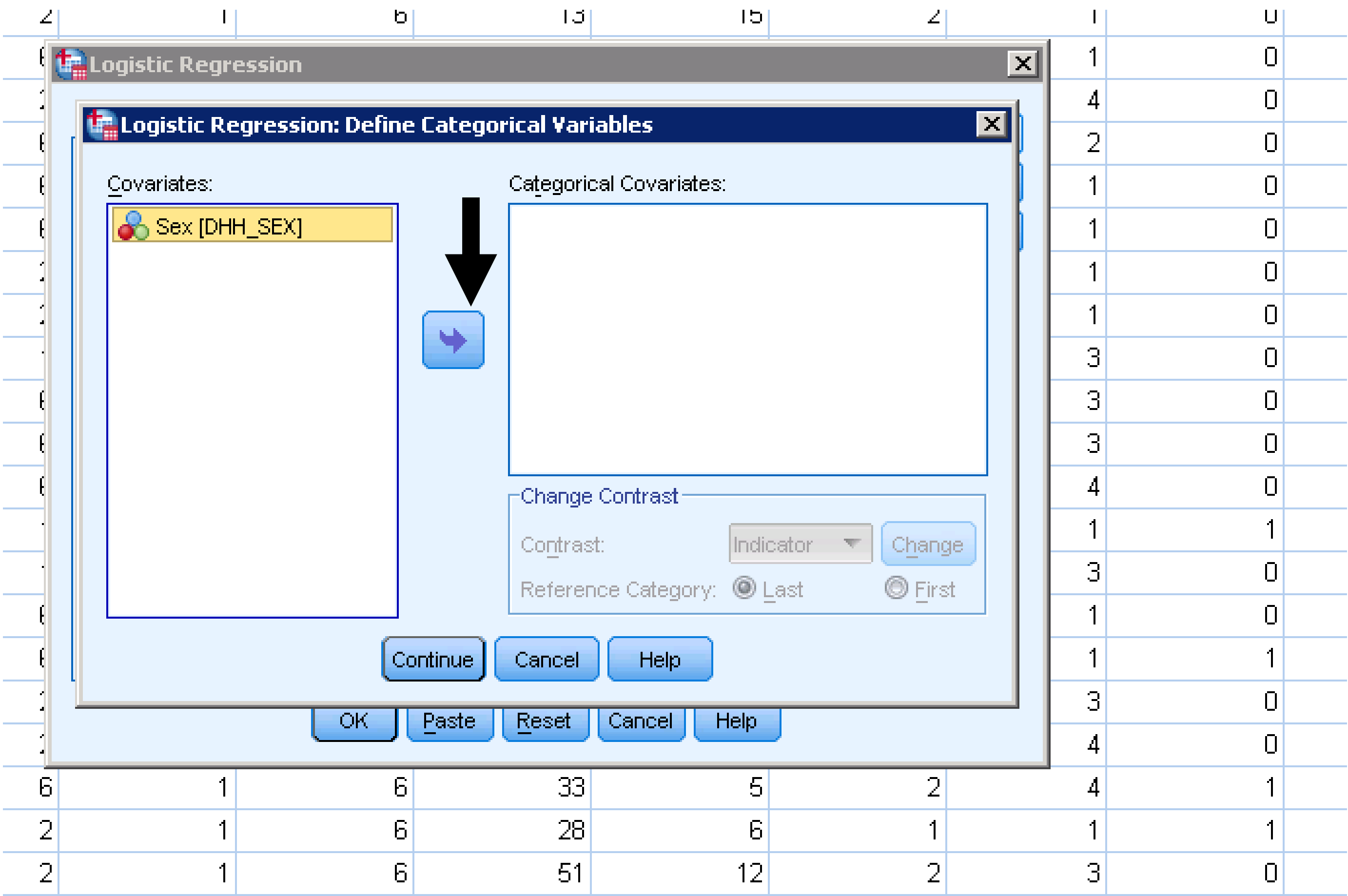


BUT:
You must click on categorical to specify "reference" Category if it isn't a "dummy variable"

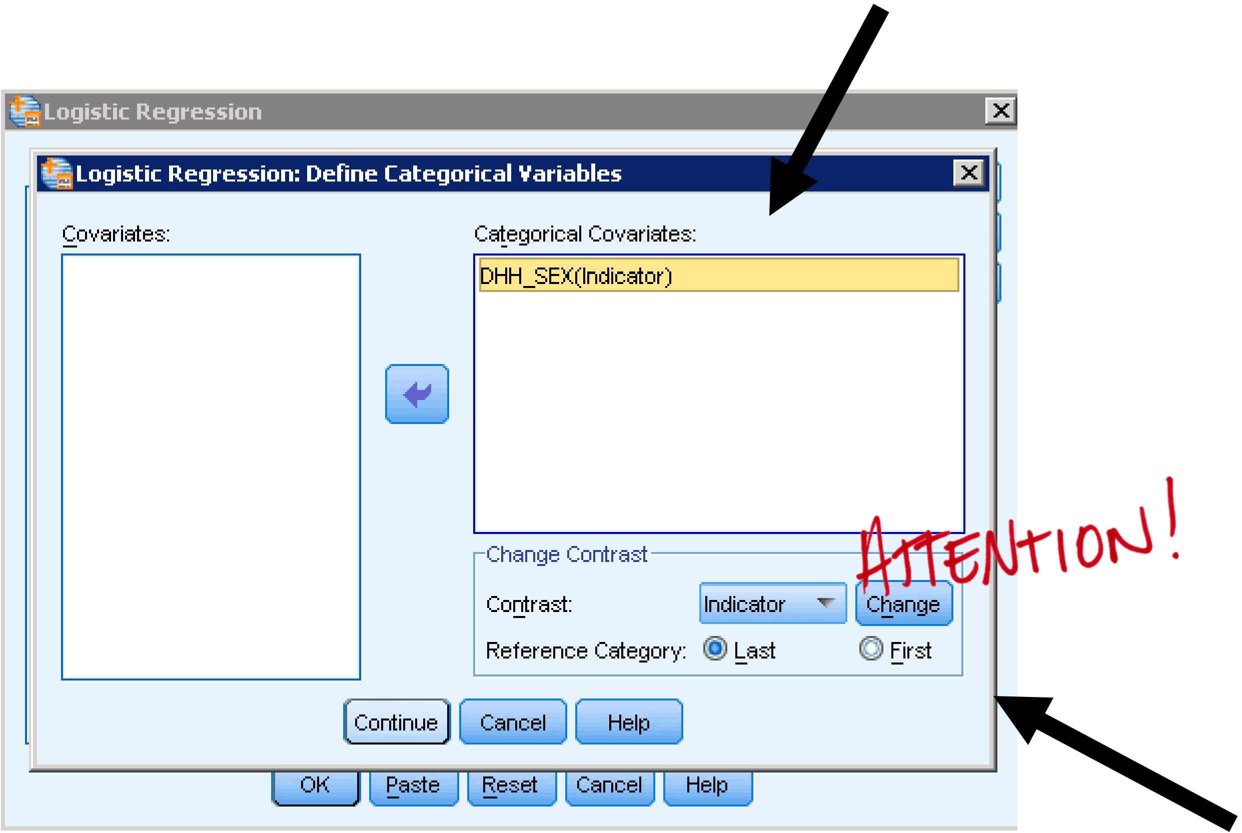
Enter the original Variable DHH_SEX

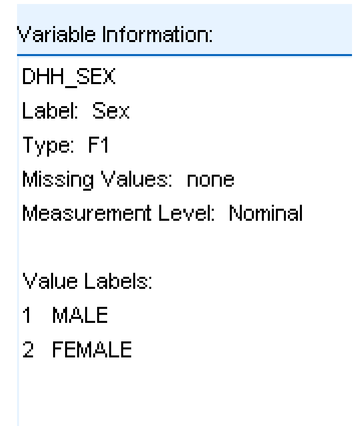


Must identify it as a "categorical covariate… click on arrow to move it over..



The variable is now identified as a “categorical” variable in the regression..





Here you must identify a reference category on DHH_SEX for our analysis; either the first or last…

Here we click “the last” to denote “FEMALE” as our reference category (don’t forget to click “change”)…

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	39722.820[a]	.003	.006

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted: Smoker .00	Predicted: Smoker 1.00	Percentage Correct
Step 1	Smoker	.00	36318	0	100.0
		1.00	7431	0	.0
	Overall Percentage				83.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DHH_SEX(1)	.309	.026	146.336	1	.000	1.361
	Constant	-1.732	.018	9313.011	1	.000	.177

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

Same result as with the dummy variable..

We denoted females as the reference category

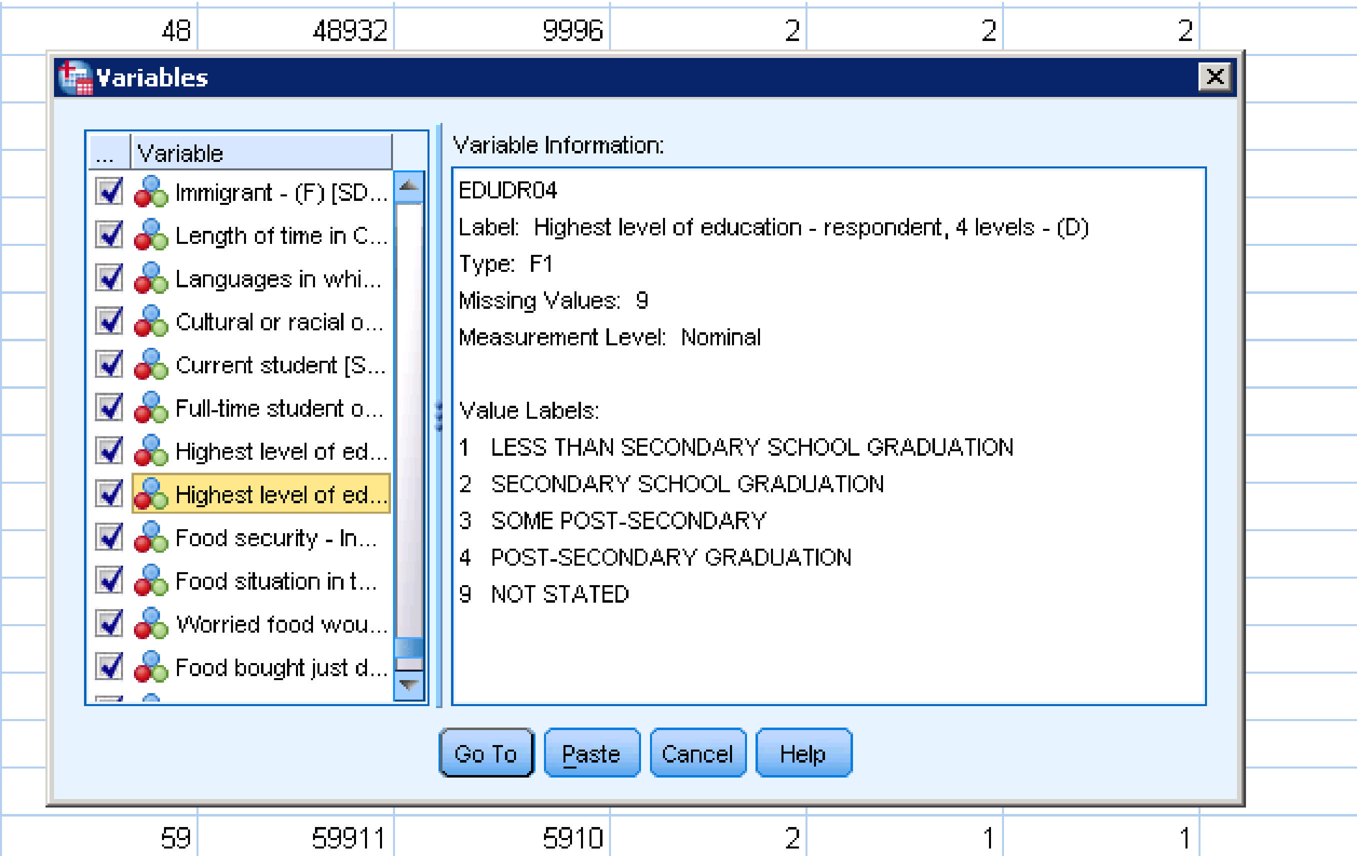
The odds are 36.1 per cent higher for males than females

Note: what if our reference category Was “male” rather than “female”?

Our Odds ratio would be: 0.659

(0.659 – 1.0)*100 -> 36.1 per cent lower

ANOTHER EXAMPLE FROM LAST WEEK:



For assignment 2 we created a dichotomous variable

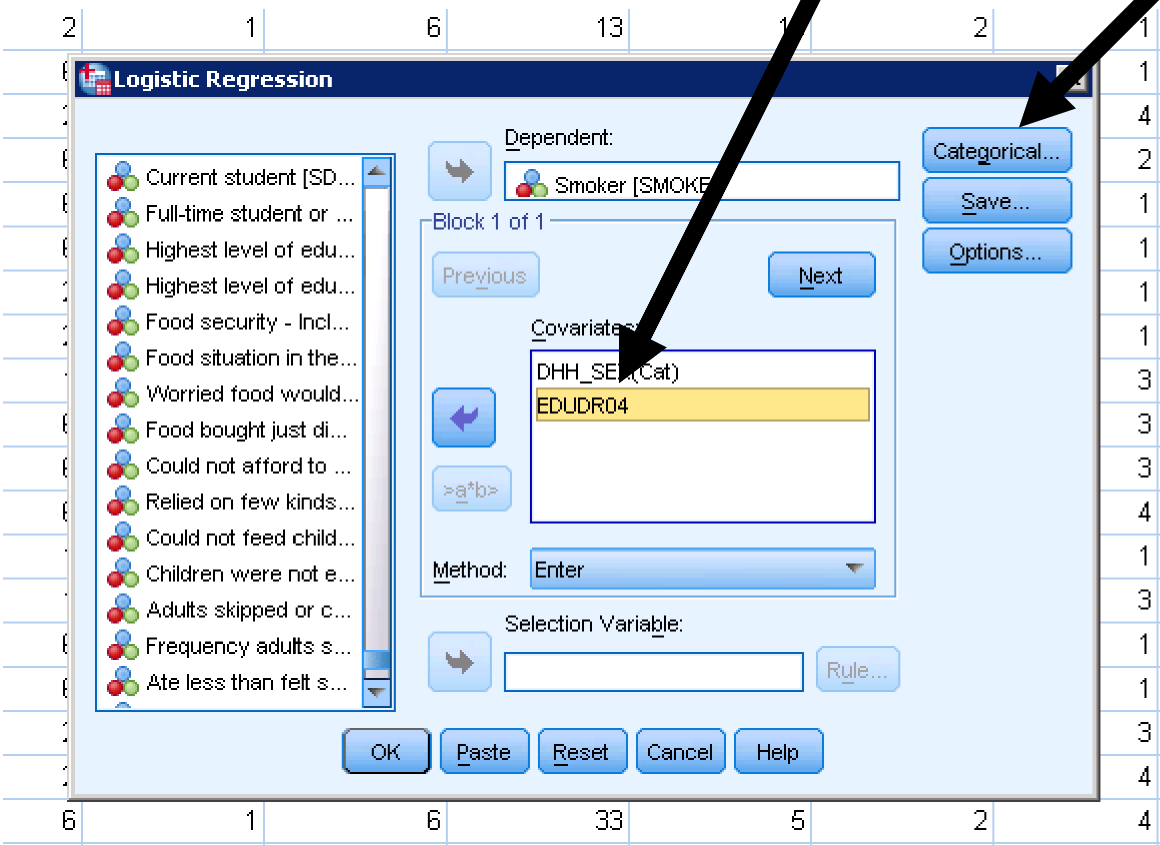
0 – not a grad

1 - grad

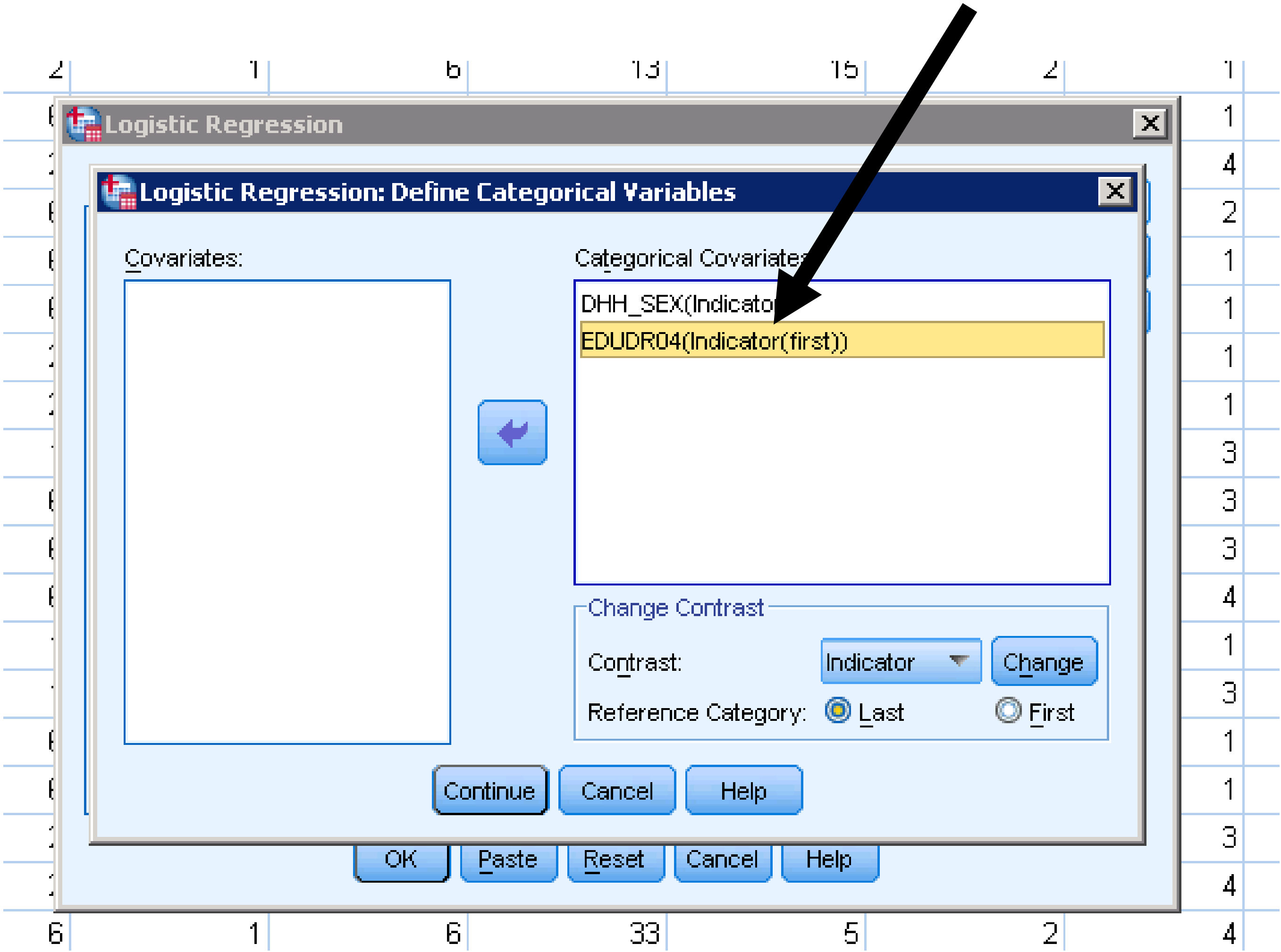
Alternatively, you can merely enter the variable as is, and correctly identify a “reference” category for our analysis..

Enter original variable..

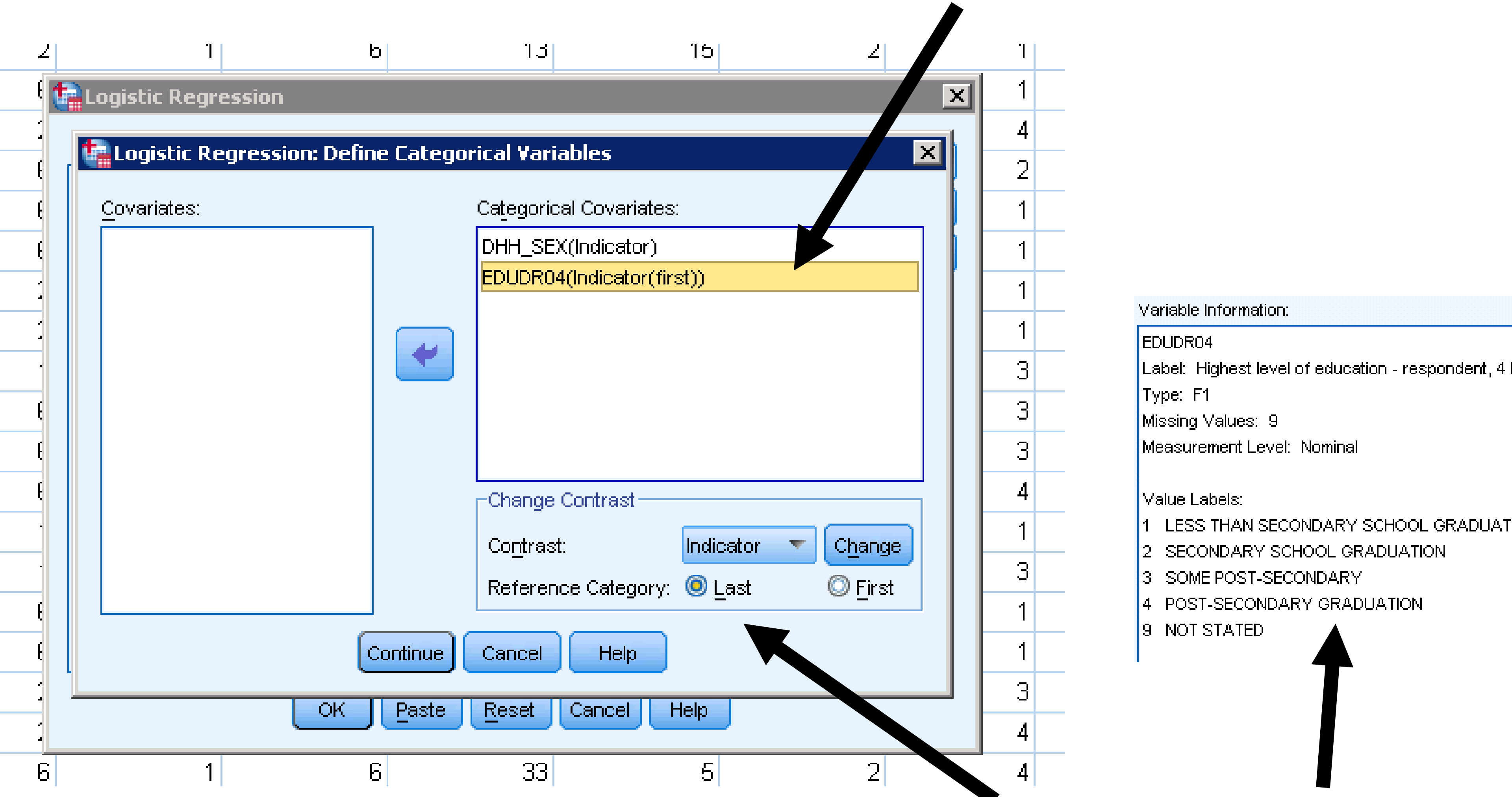
Click categorical



Assign as a categorical variable



Assign as a categorical variable



Assign “post-sec” grad
as our reference category

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	38473.238[a]	.009	.015

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

REFERENCE CATEGORY

4 POST-SECONDARY GRADUATION
9 NOT STATED

Classification Table[a]

			Predicted		
			Smoker		Percentage Correct
Observed			.00	1.00	
Step 1	Smoker	.00	35234	0	100.0
		1.00	7258	0	.0
	Overall Percentage				82.9

a. The cut value is .500

Variable Information:

EDUDR04
Label: Highest level of education - respondent, 4 levels - (D
Type: F1
Missing Values: 9
Measurement Level: Nominal

Value Labels:
1 LESS THAN SECONDARY SCHOOL GRADUATION
2 SECONDARY SCHOOL GRADUATION
3 SOME POST-SECONDARY

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DHH_SEX(1)	.318	.026	150.617	1	.000	1.375
	EDUDR04			239.735	3	.000	
	EDUDR04(1)	.418	.032	174.828	1	.000	1.518
	EDUDR04(2)	.339	.036	90.038	1	.000	1.404
	EDUDR04(3)	.463	.052	80.017	1	.000	1.589
	Constant	-1.917	.023	7097.688	1	.000	.147

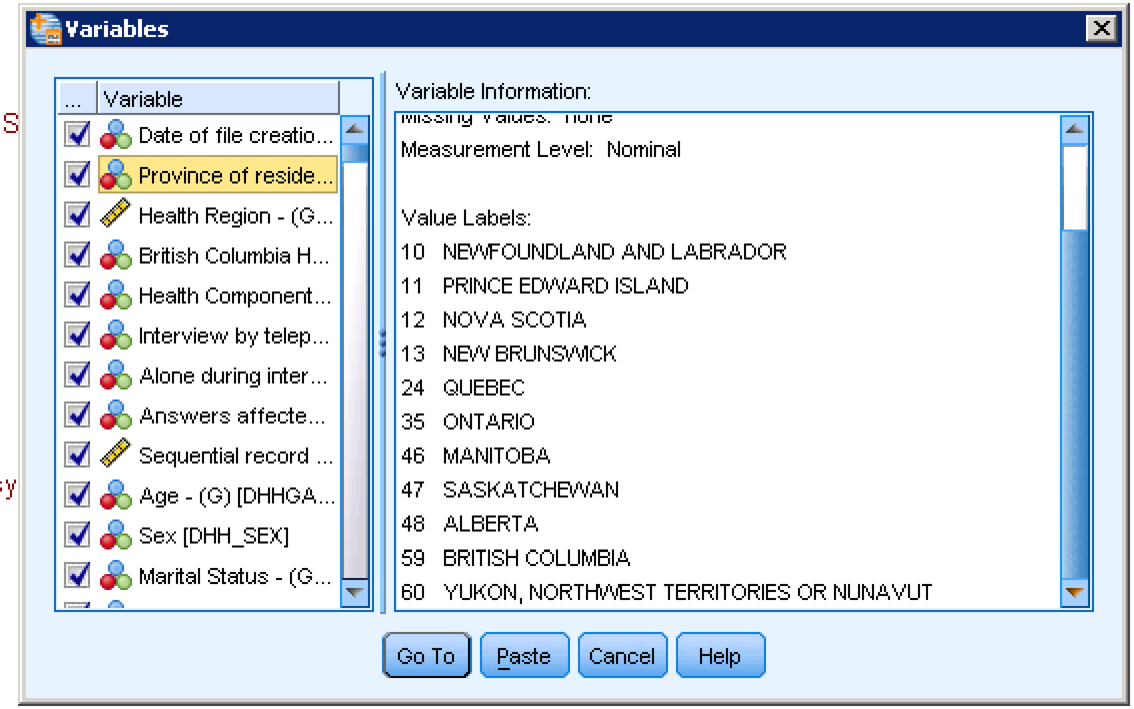
a. Variable(s) entered on step 1: DHH_SEX, EDUDR04.

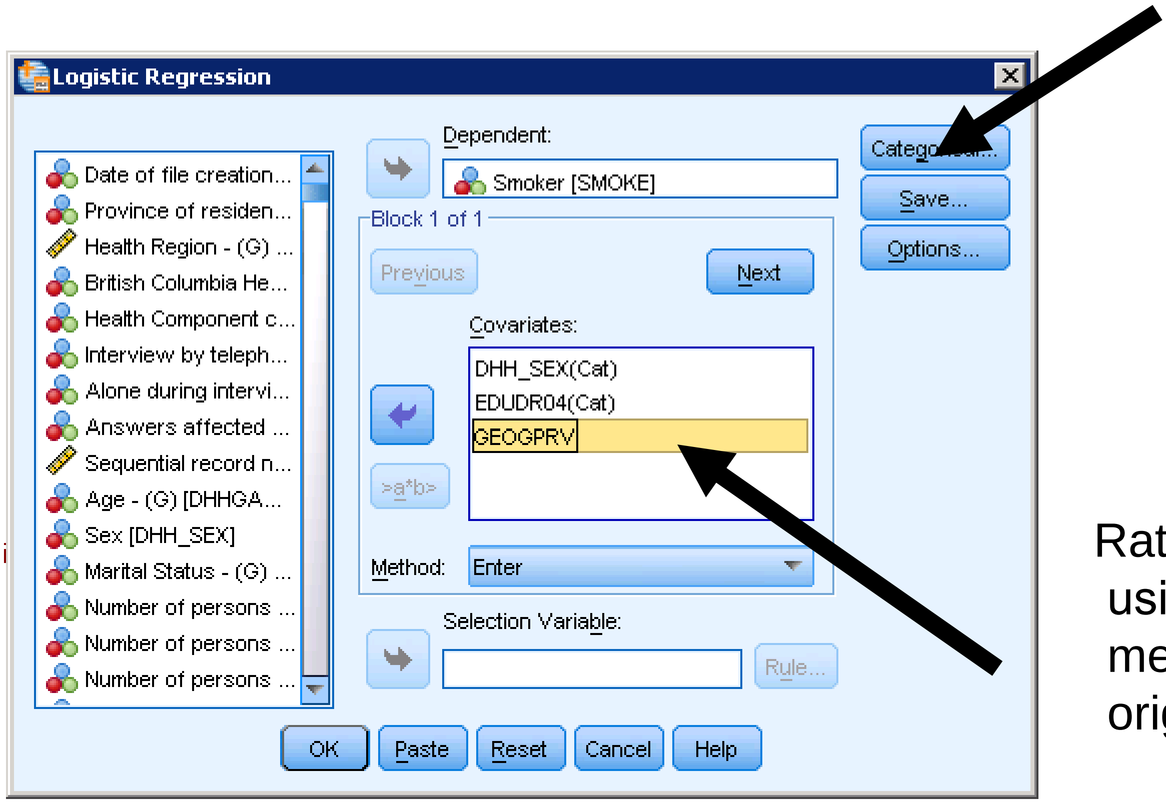
Relative to "our reference category" (post-sec grads), persons with less than secondary have 51.8 percent higher odds of smoking,

Relative to the same reference category, persons with secondary degree have 40.4 percent higher odds.

Relative to same reference, persons with "some post-secondary" have 58.9 per cent higher odds

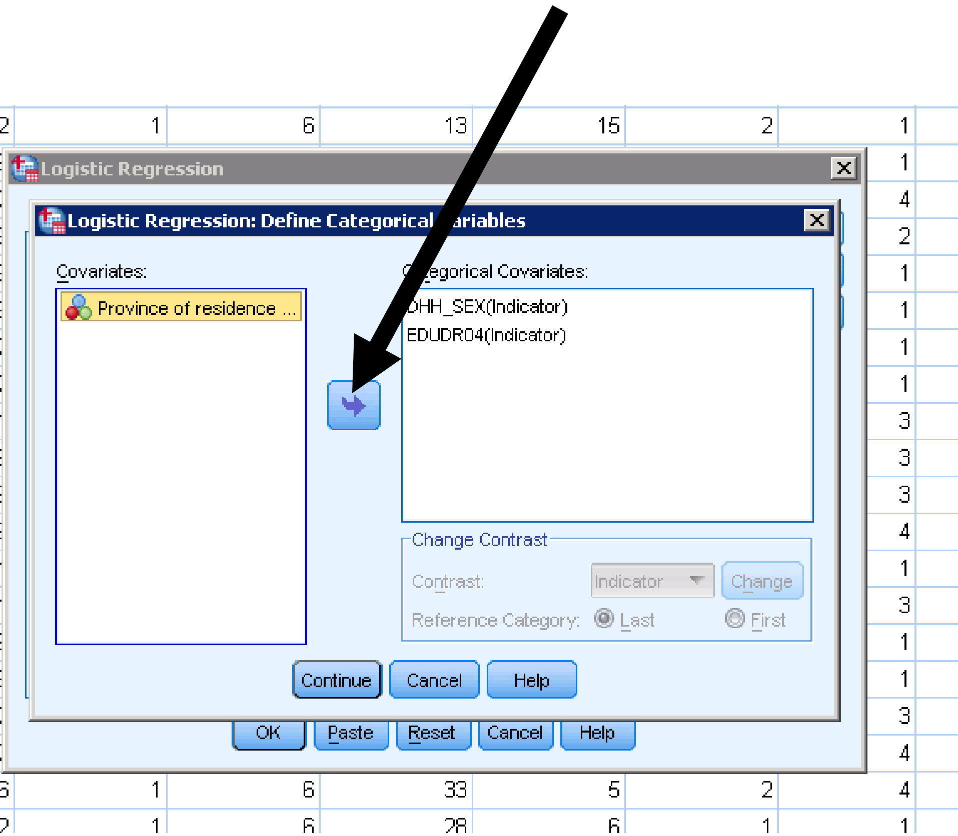
- Another example.. Say we want to consider province of residence?

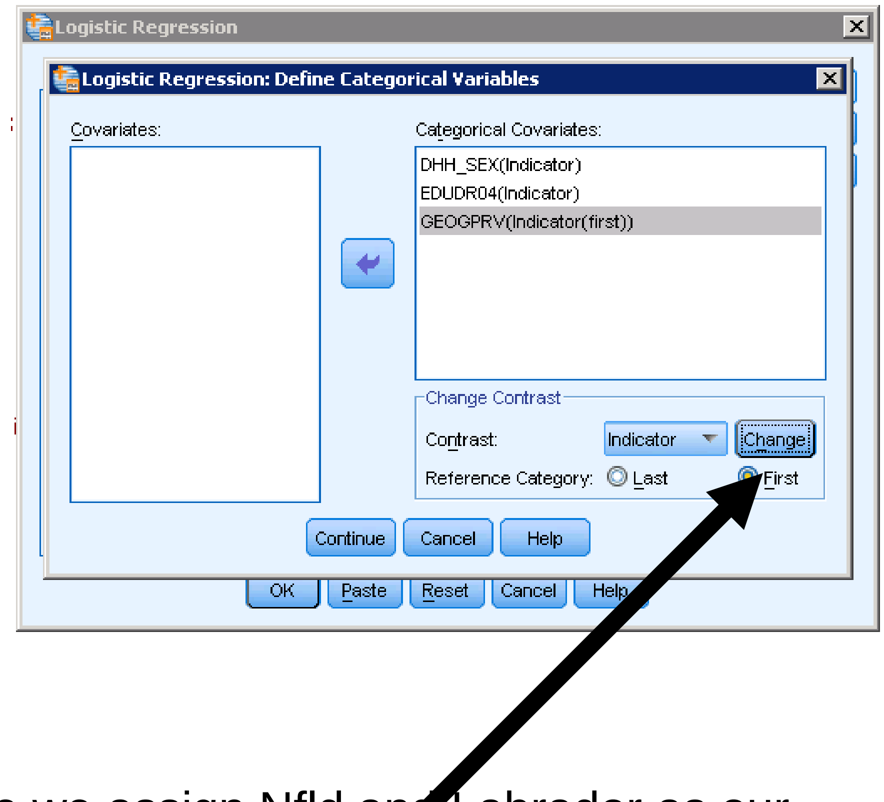




Rather than using “dummies” merely use original variable.

Assign it as a categorical variable..





Here we assign Nfld and Labrador as our reference category (the first category on GEOGPRV

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DHH_SEX(1)	.312	.026	144.062	1	.000	1.366
	EDUDR04			234.425	3	.000	
	EDUDR04(1)	.398	.032	156.017	1	.000	1.489
	EDUDR04(2)	.354	.036	97.676	1	.000	1.425
	EDUDR04(3)	.482	.052	86.201	1	.000	1.620
	GEOGPRV			169.618	10	.000	
	GEOGPRV(1)	-.026	.122	.047	1	.829	.974
	GEOGPRV(2)	.042	.095	.194	1	.660	1.043
	GEOGPRV(3)	.092	.094	.943	1	.331	1.096
	GEOGPRV(4)	.044	.077	.327	1	.567	1.045
	GEOGPRV(5)	-.153	.075	4.197	1	.040	.858
	GEOGPRV(6)	-.121	.091	1.786	1	.181	.886
	GEOGPRV(7)	-.005	.088	.003	1	.954	.995
	GEOGPRV(8)	.054	.083	.430	1	.512	1.056
	GEOGPRV(9)	-.268	.081	10.845	1	.001	.765
	GEOGPRV(10)	.666	.103	42.037	1	.000	1.946
	Constant	-1.860	.074	637.588	1	.000	.156

a. Variable(s) entered on step 1: DHH_SEX, EDUDR04, GEOGPRV.

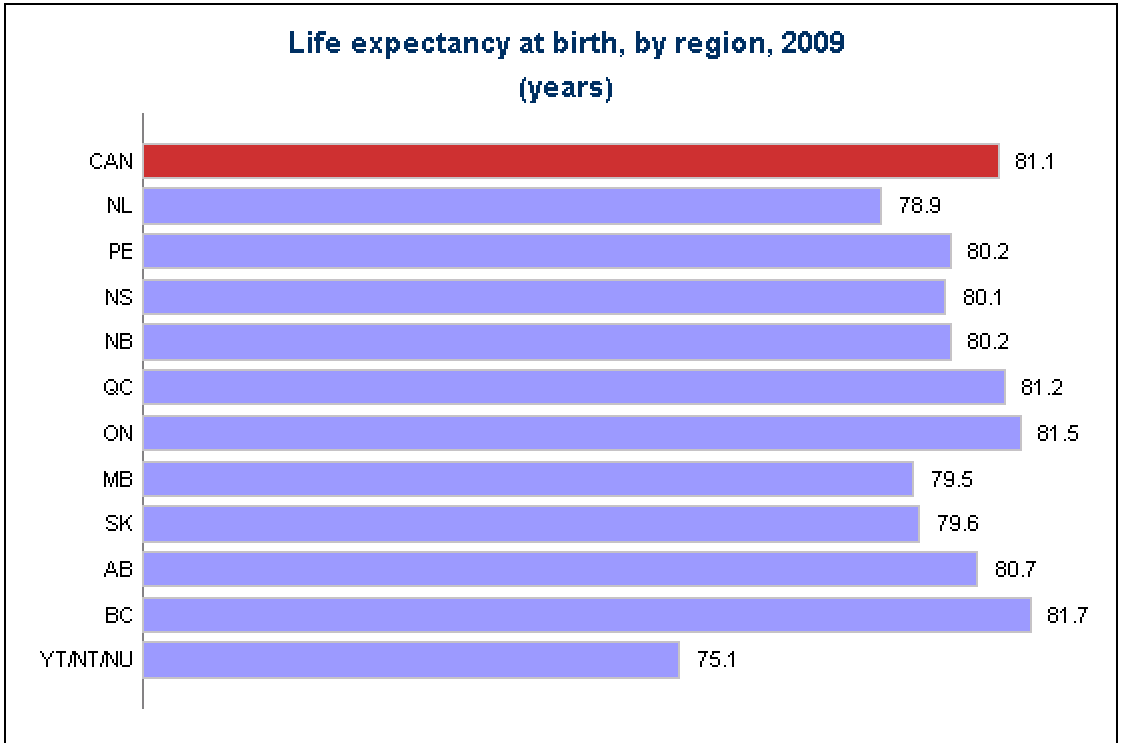
Value Labels:
10 NEWFOUNDLAND AND LABRADOR

11 PRINCE EDWARD ISLAND
12 NOVA SCOTIA
13 NEW BRUNSWICK
24 QUEBEC
35 ONTARIO
46 MANITOBA
47 SASKATCHEWAN
48 ALBERTA
59 BRITISH COLUMBIA
60 YUKON, NORTHWEST TERRITORIES OR NUNAVUT

Persons in the far north (Yukon, NWT and Nunavut) are most likely to smoke..
.. The odds are 94.6 per cent higher than in NFLD and Labr (reference)..

Persons in BC are least likely to smoke…
The odds are 23.5 per cent lower than in NFLD and Labr (reference)..
(0.765 – 1.0) * 100 = 23.5

- Substantive note:
- Did you know?



IMPORTANT REMINDER:

Again,.. this has nothing to do with OLS linear regression.

We MUST ALWAYS work with Dummy variables as independent variables when we work with nominal variables in linear regression (religion; ancestry; immigrant status, etc)..

Also:

This has nothing to do with your dependent variable in Logistic regression: We MUST always use dichotomous variables as our dependent variable (no exceptions)

- Two final things on “Logistic Regression”.. Relating to overall model performance..

- **Nagelkerke's R^2**

- **Hosmer–Lemeshow test**

- Two final things on "Logistic Regression"

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	38311.455[a]	.013	.021

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

- Nagelkerke's R^2

- In the linear regression model, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables. NOTE: THE Nagelkerke's R2 does not involve "explained variance".

- For logistic regression models with a categorical dependent variable, it is not possible to compute R^2
- *Recommendation: Use Nagelkerke's* R^2
- - referred to as a "psuedo R^2 measure"..

- Greater than 0.10 we are doing quite well… in the above example, the model is not doing a very good job in explaining our dependent variable
- $R^2 = .021$

- Technically speaking, it is based on the log likelihood for the model (all independent variables) compared to the log likelihood for a baseline model (no independent variables), adjusted to cover the full range from 0 to 1.
- (do not refer to "explained variance" with this statistic)

- **One additional test of “Goodness of Fit” indicator**
- **(indicator on overall model performance)**

- **Hosmer–Lemeshow test (we’ll consider it the “Gold” standard..)**

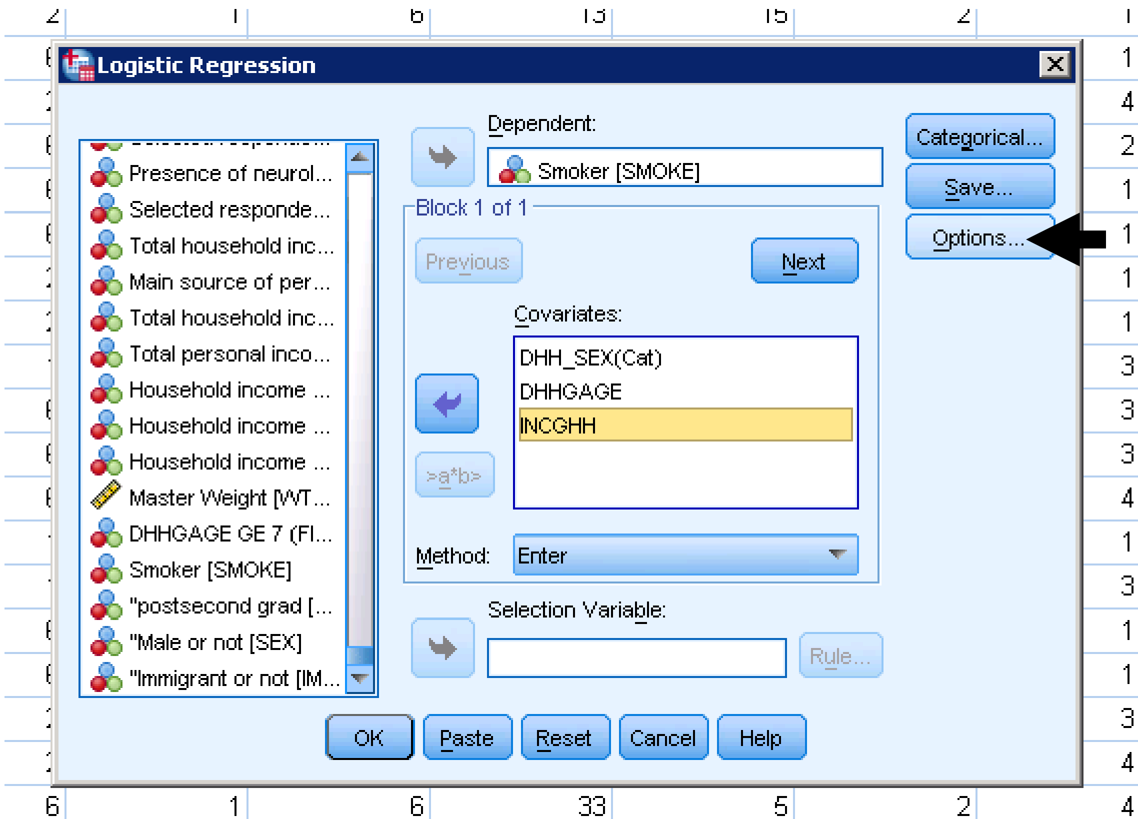
- The **Hosmer–Lemeshow test** is a statistical test for goodness of fit for logistic regression models.

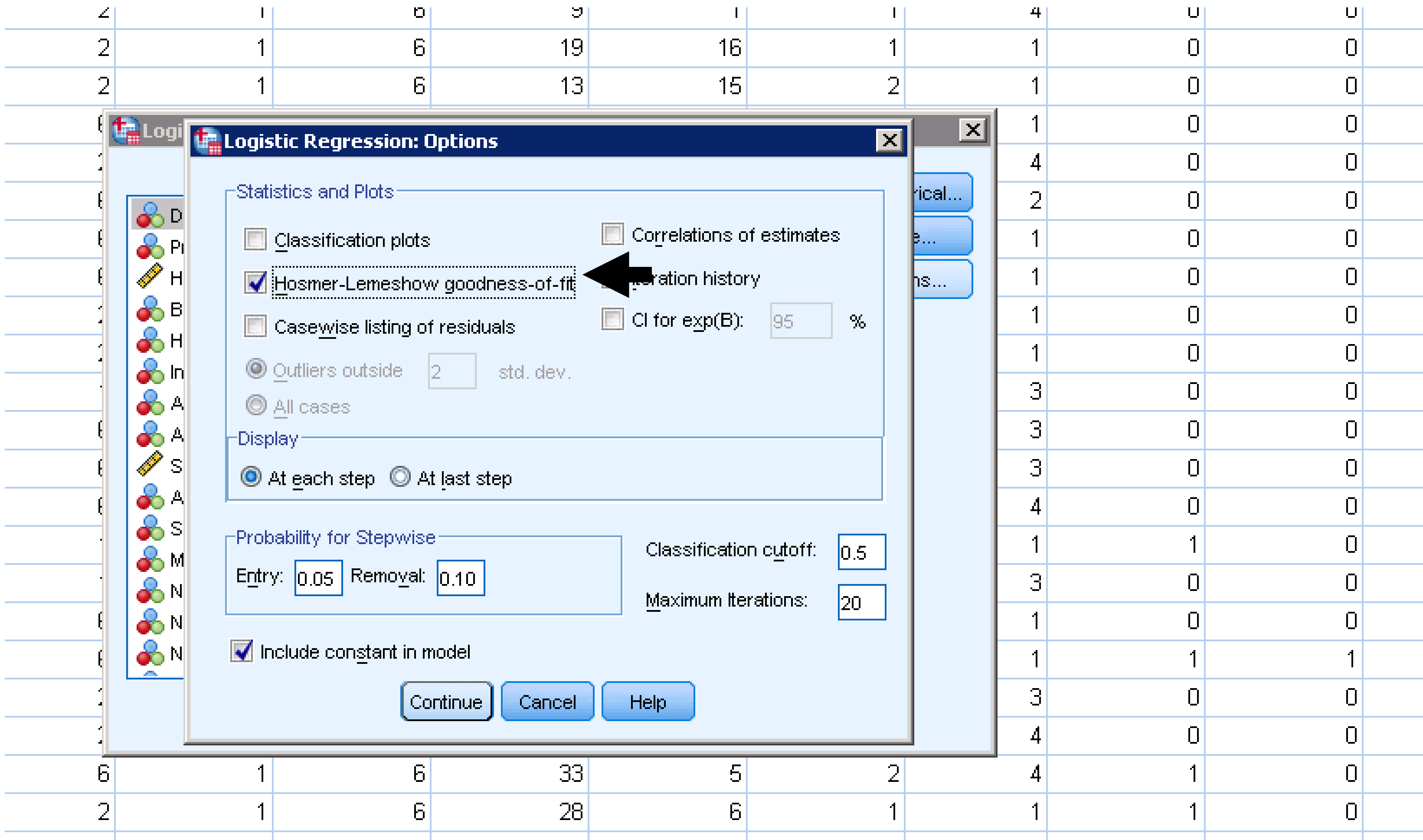
- The test assesses whether or not the observed probabilities match expected probabilities as predicted by the full model

- Recall from last week:
- Logistic regression is based on “MLE” estimation; an iterative process that attempts to come up with a series of predicted probabilities that are as close to possible to the initial observed probabilities

- This test determines helps us identify how successful MLE estimation given the variables involved..

- Goodness-of-fit tests help you decide whether your model is correctly specified (are we missing important variables?)


Logistic Regression
Presence of neurol...
Selected responde...
Total household inc...
Main source of per...
Total household inc...
Total personal inco...
Household income ...
Household income ...
Household income ...
Master Weight [WT...
DHHGAGE GE 7 (FI...
Smoker [SMOKE]
"postsecond grad [...
"Male or not [SEX]
"Immigrant or not [IM...
Dependent:
Smoker [SMOKE]
Block 1 of 1
Previous
Next
Covariates:
DHH_SEX(Cat)
DHHGAGE
INCGHH
>a*b>
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
OK
Paste
Reset
Cancel
Help


Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Display
At each step
At last step
Probability for Stepwise
Entry:
0.05
Removal:
0.10
Classification cutoff:
0.5
Maximum Iterations:
20
Include constant in model
Continue
Cancel
Help

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	38311.455[a]	.013	.021

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Counterintuitive:
In contrast to most tests of significance, here we hope for p-value > .05!!! Rather than < .05

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.829	8	.063

This is good!!

Contingency Table for Hosmer and Lemeshow Test

		Smoker = .00		Smoker = 1.00		Total
		Observed	Expected	Observed	Expected	
Step 1	1	5554	5562.029	729	720.971	6283
	2	3885	3931.091	663	616.909	4548
	3	3652	3614.828	565	602.172	4217
	4	3571	3553.663	634	651.337	4205
	5	3526	3499.153	659	685.847	4185
	6	3483	3498.249	783	767.751	4266
	7	3089	3082.477	704	710.523	3793
	8	3336	3345.276	849	839.724	4185
	9	3259	3305.669	1025	978.331	4284
	10	1879	1841.565	647	684.435	2526

We are interested in whether or not the observed probabilities match expected probabilities as predicted by the full model

Hoping for a “non-significant” difference..

- **Final comments:**
- **For the purposes of our work,.. We shall report only:**
- **Nagelkerke's R^2**

- **Hosmer–Lemeshow test can be considered a "gold standard"**
- **(we shall use it as a diagnostic tool).. But I accept a "silver" or "bronze" in this context..**

.. A p-value < .05 on this test suggests the model remains "misspecified" and that important variables have been excluded..

If you can't succeed with this., don't worry too much about it..