SOCIAL AND ECONOMIC STUDIES USING SURVEY DATA:

Part II

Examples of Analytic Use of Survey Data

Milorad S. Kovacevic

Data Analysis Research and
Data Analysis Resource Centre
Statistics Canada
kovamil@statcan.ca

Seminar for CIQSS, Montreal

February 8, 2002

4) ANALYTIC USE OF SURVEY DATA II
Categorical Data Analysis

- Consider an $a \times b$ contingency table
- Design consistent estimates of the cell proportions, p_{ij}, are given by

$$\hat{p}_{ij} = \hat{N}_{ij}/\hat{N}, \qquad \text{where} \qquad \hat{N} = \sum \hat{N}_{ij}.$$

- Now consider testing an hypothesis of no association

$$H_0: \; p_{ij} = p_{i+}p_{+j}$$

- Assumption:
  - our sample is drawn from a multinomial distribution.

- We can consider this to be a hypothesis on finite population proportions (equivalent to ***B***) or a hypothesis on model probabilities (equivalent to β).
  - in this case we will not use different symbols

Model based analysis:

uses unweighted counts, and a sample likelihood function.

- The same considerations as before apply relating to selection bias, ignorability, etc.

- The Pearson statistic for testing no association in an $a \times b$ contingency table has the form

$$X^2 = n \sum_i \sum_j \frac{(n_{ij} - n_{i+} n_{+j}/n)^2}{n_{i+} n_{+j}}$$

where n_{ij}, n_{i+}, n_{+j} and n are respectively the observed cell counts, the marginal counts and the total sample size.

- If the multinomial model is valid and H_0 is true:

$$X^2 \sim \chi^2_{(a-1)(b-1)}$$

- An analogous test statistic obtained by a simple substitution of the sample counts with the estimated population counts:

$$X_W^2 = \hat{N}\sum_i \sum_j \frac{(\hat{N}_{ij} - \hat{N}_{i+}\hat{N}_{+j}/\hat{N})^2}{\hat{N}_{i+}\hat{N}_{+j}} = \hat{N}\sum_i \sum_j \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} \quad (9)$$

- Most of the general software give this value for the test statistic for the test of independence based on the weighted counts.

- This version of test statistic, however, does not provide the meaningful measure of the deviation from the null hypothesis (Rao, Thomas, 1988).

- Instead,

$$X^2_{SW} = n \sum_i \sum_j \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} \quad . \qquad (10)$$

provides a meaningful value of the deviation from the null hypothesis.

Note (10) can be obtained in the PROC FREQ by providing scaled weights.

- However, it is DECEIVING because due to clustering the asymptotic distribution of (10) is not the $\chi^2_{(a-1)(b-1)}$ as most of analysts tend to believe.

- As a result the p-values of the test will be false unless the sampling design is accounted for properly.

Design-based analysis:

- Use weighted-up counts, and the corresponding finite population likelihood equation as the CDPQ.

- The test statistic is statistic (10) corrected for the sample design.

- One way to correct (10) for the design effect is to use the Rao-Scott corrections (Rao and Scott 1981, 1987; Thomas and Rao, 1987; Rao and Thomas, 1988).

- There are other approaches for getting a test statistic, as well **(see Thomas, Singh, Roberts, 1996)**.

- Rao-Scott first order corrections for the independence test on a two-way table:

- The first order corrected X^2 statistic is given by

$$X_c^2 = \frac{X_{SW}^2}{\hat{\delta}} \tag{11}$$

where

$$\hat{\delta} = \frac{1}{(a-1)(b-1)} \sum_i \sum_j \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{\hat{p}_{i+}\hat{p}_{+j}} \hat{d}_{ij} - \sum_i (1-\hat{p}_{i+})\hat{d}_{i+} - \sum_j (1-\hat{p}_{+j})\hat{d}_{+j}$$

- Here $\hat{d}_{ij}, \hat{d}_{i+}, \hat{d}_{+j}$ are the estimated design effects for the respective estimates $\hat{p}_{ij}$, $\hat{p}_{i+}$, and $\hat{p}_{+j}$, i.e.

$$\hat{d}_{ij} = \hat{V}(\hat{p}_{ij}) / [\hat{p}_{ij}(1-\hat{p}_{ij})/n],$$
$$\hat{d}_{i+} = \hat{V}(\hat{p}_{i+}) / [\hat{p}_{i+}(1-\hat{p}_{i+})/n],$$
$$\hat{d}_{+j} = \hat{V}(\hat{p}_{+j}) / [\hat{p}_{+j}(1-\hat{p}_{+j})/n]$$

- Knowing only the design effects for the cell and the marginal proportions we can correct X_{SW}^2 to obtain a test statistic which has the same expected value as $\chi^2_{(a-1)(b-1)}$.

- When the full estimated covariance matrix for the proportions from a two-way table is known a better approximation of the X^2 statistic is possible

  (Rao and Scott (1982, 1984))

- This approximation matches the first two moments of the corrected X^2 with the moments of χ^2 distribution, using the Satterthwaite approximation to obtain the correct number of DF.

- The second-order Rao-Scott statistic is given by

$$X_s^2 = \frac{X_{SW}^2}{\hat{\delta}\,(1+\hat{\alpha}^2)}$$

where

$$1+\hat{\alpha}^2 = \frac{1}{(a-1)(b-1)\,\delta^2}\sum_{i,i'}\sum_{j,j'}\frac{cov^2(\hat{p}_{ij}-\hat{p}_{i+}\hat{p}_{+j},\ \hat{p}_{i'j'}-\hat{p}_{i'+}\hat{p}_{+j'})}{\hat{p}_{i+}\,\hat{p}_{+j}\,\hat{p}_{i'+}\,\hat{p}_{+j'}}$$

- Second-order Rao-Scott statistic has approximately χ_ν^2 distribution where the adjusted number of degrees of freedom is $\nu=(a-1)(b-1)/(1+\hat{\alpha}^2)$.

Example

The table contains the estimated proportions obtained from the National Population Health Survey (NPHS) 1994 of the cross-classified variables on a certain health condition (2 categories) and the location of the general practitioner (8 categories):

		Location of a general practitioner							
		1	2	3	4	5	6	7	8
Health Condition	0	5.893	2.437	1.388	82.12	1.267	1.396	0.917	0.24
	1	0.221	0.208	0.162	3.593	0.044	0.118	0	0

Test statistics, DF, p-values.

Test Statistic	Value	D.F.	*p*-value
X^2_{SW}	13.538	7	0.06
X^2_c	8.585	7	0.2838
X^2_s	4.996	4.073	0.2969

p-value for X^2_{SW} is calculated on a false assumption about the χ^2 distribution

Logistic Regression

- For a binary (1 - 0) response variable Y, let

$$p_i = P(Y_i = 1 \mid \boldsymbol{x_i})$$

Then the logistic regression model is

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + , \ldots + \beta_q x_{qi},$$

or equivalently

$$p_i(\boldsymbol{\beta}) = Prob(y_i = 1 \mid \boldsymbol{x_i}, \boldsymbol{\beta}) = [1 + \exp(-\boldsymbol{x_i}' \boldsymbol{\beta})]^{-1}$$

In the non-survey context the parameter $\boldsymbol{\beta}$ is estimated by the maximum likelihood method:

- log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i \in s} \left\{ y_i \log p_i(\boldsymbol{\beta}) + (1 - y_i) \log [1 - p_i(\boldsymbol{\beta})] \right\}$$

- The score function

$$l'(\boldsymbol{\beta}) = \sum_i \boldsymbol{x}_i' [y_i - p_i(\boldsymbol{\beta})] = \sum_i u_i(\boldsymbol{\beta}) = U(\boldsymbol{\beta})$$

- The ML estimate $\hat{\boldsymbol{\beta}}$ is a solution (numeric) of the equation

$$l'(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) = 0$$

- The variance of $\hat{\beta}$ is given as the inverse of information matrix

$$\mathbf{I}(\boldsymbol{\beta}) = -l''(\boldsymbol{\beta}),$$

i.e.

$$V(\hat{\boldsymbol{\beta}}) = [\mathbf{I}(\boldsymbol{\beta})]^{-1} = [X' D(\boldsymbol{\beta}) X]^{-1}$$

Where

$$D(\boldsymbol{\beta}) = diag \left\{ p_i(\boldsymbol{\beta}) [1 - p_i(\boldsymbol{\beta})] \right\}$$

- To define the "corresponding descriptive population quantity" (CDPQ) for a logistic regression analysis we proceed as follows:

- Assume that all N finite population values are observed. The log likelihood would be

$$l_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ y_i \log p_i(\boldsymbol{\beta}) + (1 - y_i) \log [\, 1 - p_i(\boldsymbol{\beta}) \,] \right\}$$

and the score function

$$U_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{x}_i' [y_i - p_i(\boldsymbol{\beta})]$$

A solution of $U_N(\boldsymbol{\beta}) = 0$, say $\boldsymbol{B}$, is the CDPQ for a logistic regression.

- $U_N(\boldsymbol{\beta}) = 0$ are often called the "census" likelihood equations.

- Everything said for the linear regression holds for logistic as well, e.g.:
  - $\boldsymbol{B} \rightarrow \boldsymbol{\beta}$ if model is true
- The same considerations as before apply relating to selection bias, ignorability, etc.

- To estimate $\boldsymbol{B}$, the population totals in the census estimating equations are replaced by estimates, i.e., the sample estimating equations become

$$\hat{U}(\boldsymbol{B}) = 0$$

i.e.

$$\sum_{i=1}^{n} w_i y_i x_i^{/} = \sum_{i=1}^{n} w_i p_i(\boldsymbol{B}) x_i^{/},$$

which must be solved iteratively for $\hat{\boldsymbol{B}}$.

- This is a very general procedure that can be applied to a variety of generalized linear models. It is referred to as the "**pseudo (weighted) likelihood**".

- Variance estimator for $\hat{\boldsymbol{B}}$ may be obtained by the Taylor linearization method for implicitly-defined parameters (Binder, 1983):

  ("Sandwich estimator")

$$\hat{V}_L(\hat{\boldsymbol{B}}) = (\boldsymbol{J}^{-1})\ \hat{V}_L[\hat{U}(\hat{\boldsymbol{B}})]\ (\boldsymbol{J}^{-1})$$

  where $\boldsymbol{J}$ is the weighted sample information matrix

$$\boldsymbol{J} = X_n^{/} W_n D(\hat{\boldsymbol{B}}) X_n$$

  and the score functions are simple linear functions of data:

$$\hat{U}(\hat{\boldsymbol{B}}) = \sum_i w_i \boldsymbol{x}_i^{/} (y_i - \hat{p}_i)$$

Logistic regression example

Example motivated from "The Upward Mobility of Low Paid Canadians" by Drolet and Morissette (1998))

Population of interest

Paid workers who:

- were aged 15-60 in 1993
- were not enrolled in school full-time in 1993 or 1995
- were employed in December 1993 and December 1995
- had low weekly earnings (in their main job) in 1993.

Low weekly earnings threshold

In 1993, weekly earnings < \$404.16 = low earnings.
In 1995, weekly earnings < \$413.86 = low earnings.

Individuals having weekly earnings > \$455.25 in 1995 are said to have moved out of low earnings.

OBJECTIVE: To study the upward mobility of these 1993 low weekly earners between 1993 and 1995

Number of observations in each regression = 2,188.
Using data from the Survey of Labour and Income Dynamics (SLID)

TABLE: Odds ratios from logistic regressions predicting the probability of moving out of low earnings, 1993-95, using assorted software and methods

INDEPENDENT VARIABLES	CATEGORY		SAS		SUDAAN		WES VAR	IN-HOUSE
		NO WGT	UNSCLD	SCLD	REPLIC	Taylor	JK	JK
AGE IN 1993	15-24	.76	.68**	.68*	.68	.68	.68	.68
	25-34	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
	35-44	0.95	1.07**	1.07	1.07	1.07	1.07	1.07
	45-54	1.22	1.10**	1.1	1.1	1.1	1.1	1.1
	55-60	.41*	.19**	.19**	.19**	.19**	.19**	.19**
FAMILY COMPOSITION & GENDER IN 1993	Male	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
	Female married, no children	.29**	.26**	.26**	.26**	.26**	.26**	.26**
	Female married, with children	.28**	.28**	.28**	.28**	.28**	.28**	.28**
	Female, lone parent	.19**	.19**	.19**	.19**	.19**	.19**	.19**
	Female single, never married	.48**	.72**	.72	.72	.72	.72	.72
	Other, no children	.20**	.14**	.14**	.14**	.14**	.14*	.14**
EDUCATION	Secondary or less	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
	Some post secondary	1.42*	1.14**	1.14	1.14	1.14	1.14	1.14
	College/univ degree	2.18**	1.59**	1.59*	1.59	1.59	1.59	1.59
OCCUPATION	Prof. Management & sciences	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
	Clerical	.60**	.60**	.60**	.6	.6	.6	.6
	Sales	.48*	.37**	.37**	.37**	.37**	.37**	.37**
	Services	.36**	.33**	.33**	.33**	.33**	.33**	.33**
	Blue collar	.48**	.45**	.45**	.45*	.45*	.45*	.45*
REGION	Atlantic Provinces	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
	Québec	1.66*	1.88**	1.88**	1.88**	1.88**	1.88*	1.88**
	Ontario	2.46**	2.66**	2.66**	2.66**	2.66**	2.66**	2.66**
	Pairie Provinces	1.18	1	1	1	1	1	1
	Alberta	1.97**	2.02**	2.02*	2.02*	2.02*	2.02*	2.02*
	British Columbia	3.28**	2.89**	2.89**	2.89**	2.89**	2.89**	2.89**

INDEPENDENT VARIABLES	CATEGORY		SAS		SUDAAN		WES VAR JK	IN-HOUSE JK
		NO WGT	UNSCLD	SCLD	REPLIC	Taylor		
INDUSTRY	Goods producing sector	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**
	Distributive & Bus. Services	**.62***	**.87****	**.87**	**.87**	**.87**	**.87**	**.87**
	Consumer services	**.24****	**.35****	**.35****	**.35****	**.35****	**.35****	**.35****
	Public services	**.51****	**.79****	**.79**	**.79**	**.79**	**.79**	**.79**
CHANGED JOBS	Yes	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**
	No	**.82**	**.83****	**.83**	**.83**	**.83**	**.83**	**.83**
CHANGED UNION STATUS	Non-union 93➔ union 95	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**
	Other	**.77**	**.75****	**.75**	**.75**	**.75**	**.75**	**.75**
CHANGE IN FIRM SIZE	Small firm 93➔ large firm 95	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**
	Other	**0.62**	**.61****	**.61**	**.61**	**.61**	**.61**	**.61**
NUMBER OF SIGNIFICANT TESTS OUT OF 26		17	25	16	13	13	13	13

Reference categories, whose odds ratios, which are identically 1.00, are in parentheses.

* Significant at level .05 (but not at level .01)

** Significant at level .01

Details of logistic regression results for College/university degree, using assorted software and methods

	SAS			SUDAAN		WES VAR	IN HOUSE
	NO WGTS	UNSCLD	SCLD	REPLICATE	TAYLOR	JK	JK
Coefficient	0.7779	0.4623	0.4623	0.4623	0.4623	0.4623	0.4623
Standard error of coefficient	0.2544	0.0081	0.2216	0.3254	0.3236	0.3254	0.3272
Odds ratio	2.18**	1.59**	1.59*	1.59	1.59	1.59	1.59
95% interval for odds ratio	**(1.32, 3.57)**	**(1.56, 1.61)**	**(1.02, 2.44)**	**(.84, 3.01)**	**(.84, 3.00)**	**(.84, 3.01)**	**(.83,3.02)**
P value	0.0022	0.0001	0.037	0.156	0.1544	0.156	0.1583

* Significant at level .05 (but not at level .01)
** Significant at level .01

- Describes the hazard function as a non-linear function of a linear combination of predictor variables that may be functions of time, i.e.,

$$h(t|\boldsymbol{x}(t)) = h_0(t)\exp[\boldsymbol{x}(t)'\beta]$$

where $h_0(t)$ is the unknown baseline hazard.

- The hazard function can be thought of as the instantaneous probability of an event (death, losing a job, etc.) occurring at time *t*, given that the event has not occurred prior to *t*, i.e.,

$$h(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} Prob(t \le T \le t + \Delta \,|\, T > t) = \frac{f(t)}{1 - F(t)}$$

where $f(t)$ and $F(t)$ are the density and cumulative distribution, respectively, for the random variable *T*.

- Thus assessing the effect of potential risk factors, the $x_j(t)$, on the hazard becomes an analytically powerful tool for use in longitudinal studies.

- The same problems with the model (as before) can occur:
  - missing covariates / predictors
  - misspecification of the predictor such as non-linearity
- In addition, when the data come from a complex survey
  - non-independence of observations due to clustering
  - design variables not included in the model

- Design based estimation can again provide estimators of β that protect against misspecifications and non-ignorability of the design (Binder, 1992)

- Binder (1992) defined a CDPQ $\boldsymbol{B}$ of β using the partial likelihood customarily optimized for i.i.d. samples.

  – thus point estimates $\hat{\boldsymbol{B}}$ were obtained by solving weighted sample estimating equations

  – variance estimation required a modification of the Binder (1983) approach, to account for the time dependence.

Example:
JOBLESS SPELLS ANALYSIS

Motivating study:

Galarneau, D. and Stratychuk, L.M. (2001) After the Layoff. Perspectives on Labour and Income. Statistics Canada

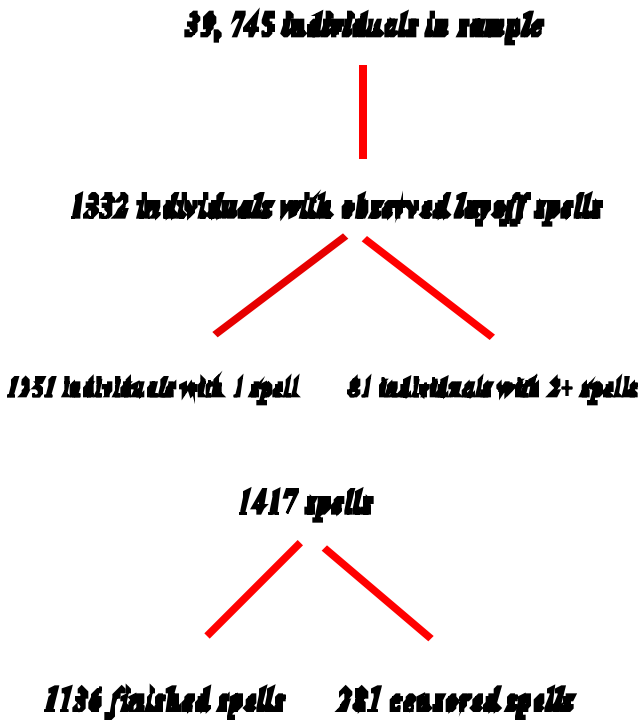
Source of Data: SLID (Panel I)
6 years of data (1993 to 1998)

To be studied:
Factors related to duration of jobless spells in the Canadian population in the 1990's

Type of spell to be studied:
Time period without employment after permanent layoff from a full-time job where:

- job had lasted at least 12 months,
- layoff was for certain reasons,
- the person holding the job had no other jobs,
- the person holding the job was not full-time student

Description of Data for Study



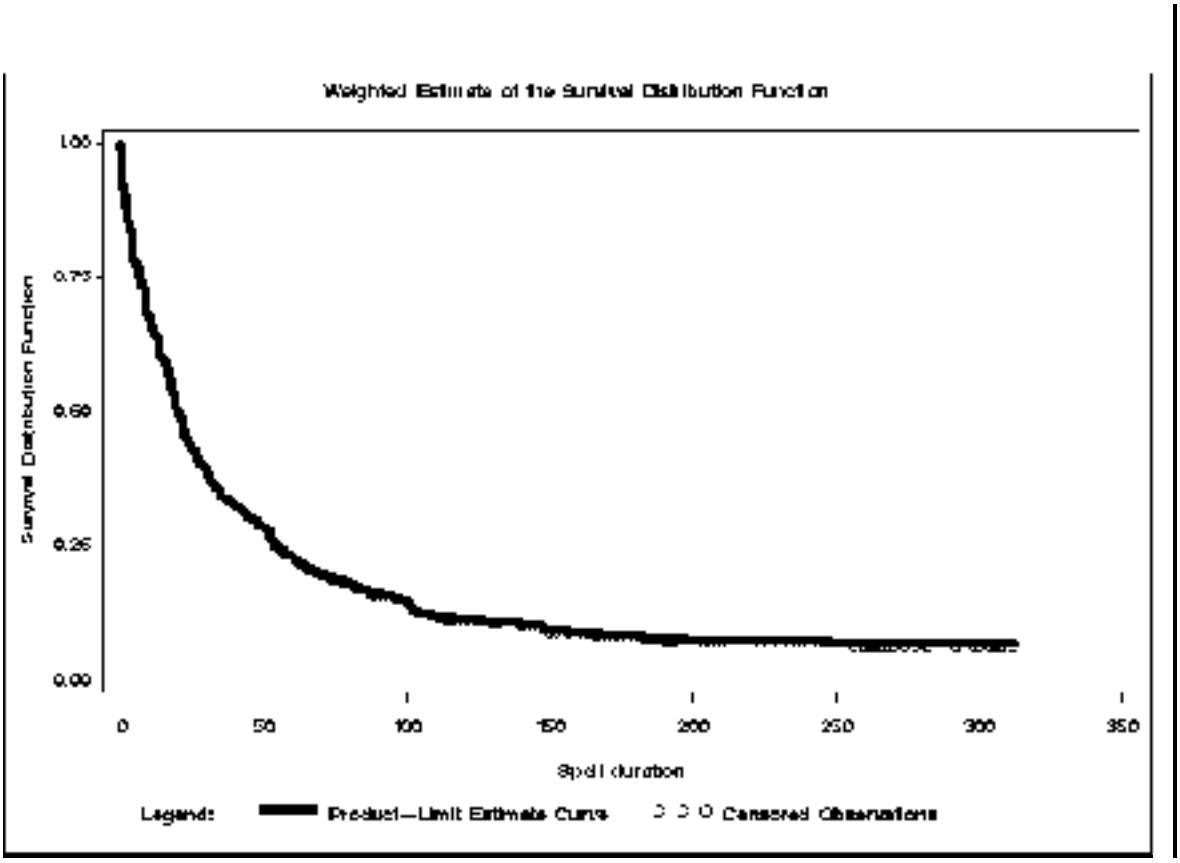
Factors proposed by researcher to be considered:

- Person characteristics
- Family or household characteristics
- Layoff job characteristics
- Timing of the spell
- Person-spell characteristics

Estimation of Survival function

Let $S(t) = Prob(event\ is\ longer\ than\ t)$

Survival function can be estimated using SAS PROC Lifetest-with deficiencies:truncation of weights to integers, no variance estimates



Variable	SAS-UNWEIGHTED		SAS-WEIGHTED	p-values		SUDAAN	
	BETA	p-values		UNSCALED	SCALED	BETA	p-values
AGESEX							
m1625	0	.	0	.		0	.
m2535	0.12058	0.4322	0.17377	<.0001	0.2029	0.1738	0.3602
m3545	0.04214	0.7933	-0.00459	0.3733	0.9751	-0.0046	0.9826
m4555	-0.06883	0.6815	-0.03382	<.0001	0.8241	-0.0338	0.8765
m55p	**-0.93842**	<.0001	**-1.14587**	<.0001	<.0001	**-1.1459**	0.0001
fl625	-0.02749	0.8993	0.09058	<.0001	0.6481	0.0906	0.7673
f2535	-0.16541	0.3387	-0.28010	<.0001	0.0765	-0.2801	0.2961
f3545	-0.19091	0.2678	-0.24397	<.0001	0.1253	-0.2440	0.3397
f4555	-0.30421	0.1083	-0.10289	<.0001	0.5372	-0.1029	0.6601
f55p	**-1.12638**	<.0001	**-1.42618**	<.0001	<.0001	**-1.4261**	0.0000
EDUC							
low_edu	-0.12797	0.1452	**-0.18083**	<.0001	0.036	-0.1808	0.1044
mid_educ	0	.	0	.		0	.
hi_educ	0.06731	0.3821	-0.02970	<.0001	0.6848	-0.0297	0.7357
VISMIN							
No	0.26809	0.1073	**0.31397**	<.0001	0.0095	**0.3140**	0.0252
SPOUSAL INCOME							
rc0000	0.05565	0.6423	0.17243	<.0001	0.1315	0.1724	0.1867
rc0010	0.13003	0.3155	0.09721	<.0001	0.4453	0.0972	0.5378
rc1020	0.00475	0.9708	-0.10617	<.0001	0.4165	-0.1062	0.5526
rc2040	0.04086	0.7262	-0.11324	<.0001	0.3235	-0.1132	0.3790
rc40+	0	.	0	.	.	0	.
PROFESSION							
prim	0.16666	0.3418	0.08400	<.0001	0.6756	0.0840	0.1505
constr	**0.30462**	0.0203	0.10832	<.0001	0.4678	0.1083	0.1830
pro	**0.28164**	0.0018	**0.41819**	<.0001	<.0001	**0.4182**	0.0008
cvserv	0	.	0	.	.	0	.
process	0.14836	0.1184	**0.21711**	<.0001	0.0133	0.2171	0.0904
other	0.15026	0.2427	**0.31153**	<.0001	0.0106	0.3115	0.6314
MONTH							
jan	0.11768	0.4753	-0.15098	<.0001	0.3606	-0.1510	0.5328
feb	0.00997	0.9533	-0.06917	<.0001	0.6896	-0.0692	0.7237
march	0	.	0	.		0	.
apr	0.01429	0.9313	-0.22658	<.0001	0.1589	-0.2266	0.3294
may	0.12740	0.4284	0.17520	<.0001	0.2387	0.1752	0.2974
jun	-0.01122	0.9423	-0.21748	<.0001	0.1598	-0.2175	0.1988
jul	0.03132	0.8535	-0.12087	<.0001	0.4050	-0.1209	0.4222
aug	0.20575	0.2322	0.18343	<.0001	0.2559	0.1834	0.3414

Variables	SAS-UNWEIGHTED		SAS-WEIGHTED	p-values		SUDAAN	
	BETA	p-values		UNSCALED	SCALED	BETA	p-values
sep	0.05618	0.7216	0.00134	0.8092	0.9932	0.0013	0.9952
oct	0.10543	0.5002	0.00140	0.7820	0.9922	0.0014	0.9759
nov	0.12039	0.4344	0.03579	<.0001	0.8050	0.0358	0.8629
dec	**0.36682**	0.0114	0.19406	<.0001	0.1731	0.1941	0.2987
HOURLY WAGE							
sali0007	**0.22138**	0.0422	0.09059	<.0001	0.4259	0.0906	0.5690
sali0710	0	.	0	.	.	0	.
sali1015	0.11072	0.2194	-0.01230	0.0001	0.8915	-0.0123	0.8902
sali15p	0.07595	0.4174	0.02287	<.0001	0.8031	0.0229	0.8959
REC_EI							
Yes	**-0.35045**	<.0001	**-0.49054**	<.0001	<.0001	**-0.4905**	0.0000
Duration of the laid-off job							
more than 60m	-0.08016	0.2656	**-0.15422**	<.0001	0.0322	-0.1542	0.1027
REGION							
atl	0.11324	0.2451	0.01880	<.0001	0.8811	0.0188	0.5123
pq	-0.06216	0.4903	-0.10929	<.0001	0.1545	-0.1093	0.3025
ont	0	.	0	.	.	0	.
prairies	**0.38456**	0.0007	**0.32859**	<.0001	0.0171	**0.3286**	0.0430
alb	0.19386	0.0712	0.08462	<.0001	0.4129	0.0846	0.6228
bc	0.23319	0.0661	0.15946	<.0001	0.1562	0.1595	0.3759
YEAR of incidence							
a93	0	.	0	.	.	0	.
a94	0.12136	0.2575	0.13321	<.0001	0.1777	0.1332	0.2822
a95	**0.22758**	0.0274	**0.24986**	<.0001	0.0105	0.2499	0.0557
a96	**0.26018**	0.0158	**0.40007**	<.0001	0.0001	**0.4001**	0.0244
a97	**0.50603**	<.0001	**0.58237**	<.0001	<.0001	**0.5824**	0.0001
a98	-0.02556	0.8551	-0.26355	<.0001	0.0659	-0.2635	0.1665
# of significant BETAs	11			44	12	8	

SUMMARY AGAIN:

	Assumed model is valid	Model is misspecified
Model-based	- Consistent - Efficient - Valid variance estimates - Valid inference - May be best	- May be inconsistent - Variance estimates may be invalid - Inference may be invalid
Design-based	- Consistent - May be inefficient - Valid variance estimates - Valid inferences	- Consistent for model parameter - Valid conditional variance estimates - valid estimates of total variance - valid inference

ANALYTICAL METHODS ACCOMMODATED

	SAS SVY	Stata 6	SUDAAN 7.5/8	WesVar 3.0 /4
Means, proportions, totals, ratios; associated se's	Y	Y	Y	Y
Quantiles			Y	Y
Tests of Independence		Y	Y(-)	Y
Contrasts	Y	Y	Y	Y
Linear regression	Y	Y	Y	Y
Logistic regression		Y	Y	Y
Multinomial logistic - ordinal and nominal categories		Y	Y	
Proportional hazards model			Y	
Probit model		Y		
Instrumental variables regression, censored and interval regression, Poisson regression		Y		
Graphical diagnostics	Y(-)	Y(-)		