

End Term

January, 2009

Take the 5% two-tailed critical value of the normal distribution to be 2. For t , F and chi-squared tests report the statistic and the corresponding degrees of freedom. There's no need to calculate P -values.

[1] Cigarette Smoking (30 pts)

Wooldridge (2002) has an interesting dataset on cigarette smoking. The outcome of interest is *cigs*, the number of cigarettes smoked per day. Predictors include *lcigpric*, the log of the price of cigarettes in the state (cents/pack); *lincome*, the log of income; *restaurn*, a dummy variable indicating whether there are restaurant smoking restrictions in the state; *white*, a dummy variable for whites; *educ*, the number of years of education, and *age* and *agesq* representing linear and quadratic terms on age.

- A Poisson regression with all predictors gives a log-likelihood of -8111.52 and a Pearson chi-squared statistic of $16,232.71$. The estimated coefficient of *restaurn*, with standard error in parentheses, is -0.3636 (0.0312).
- A negative binomial model with the same predictors yields the results shown below

```
Negative binomial regression          Number of obs   =          807
                                     LR chi2(7)      =          22.53
Dispersion   = mean                 Prob > chi2     =          0.0021
Log likelihood = -1929.8502          Pseudo R2      =          0.0058
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
<i>cigs</i>						
<i>lcigpric</i>	-.2040058	1.133861	-0.18	0.857	-2.426333	2.018322
<i>lincome</i>	.1024636	.1534655	0.67	0.504	-.1983232	.4032504
<i>restaurn</i>	-.4704123	.2307134	-2.04	0.041	-.9226023	-.0182222
<i>white</i>	-.1823308	.3082607	-0.59	0.554	-.7865107	.4218491
<i>educ</i>	-.0944912	.0401615	-2.35	0.019	-.1732062	-.0157761
<i>age</i>	.1360284	.0342455	3.97	0.000	.0689083	.2031484
<i>agesq</i>	-.0016187	.0003745	-4.32	0.000	-.0023528	-.0008847
<i>_cons</i>	.9634568	4.614216	0.21	0.835	-8.080241	10.00715
<i>/lnalpha</i>	1.997262	.0690029			1.862018	2.132505
<i>alpha</i>	7.36885	.5084723			6.436716	8.435972

- Concerned that non-smokers may inflate the frequency of zeroes, an analyst fits a zero-inflated Poisson model, getting a log-likelihood of -2349.38 by adding one parameter to account for extra zeroes (a constant in the inflate equation). Including all predictors in the inflate equation increases the log-likelihood to -2322.07 , but we will not consider that model.

(a) Interpret the coefficient of *restaurn* in the Poisson model and test its significance.

Cigarette consumption is 30.5% lower in states with restaurant smoking restrictions than in those without such restrictions, after adjusting for the price of cigarettes and respondent's income, education, ethnicity and age [$\exp(-0.3636) - 1 = -0.3048$]. The coefficient is highly significant ($z = -0.3636/0.0312 = -11.65$).

- (b) Is there evidence of over-dispersion? Assume the variance is proportional rather than equal to the mean and estimate the proportionality factor. Use this information to revisit the question of whether the effect of restaurant restrictions is significant after allowing for Poisson over-dispersion.

Yes, there is clear evidence of over-dispersion. Using Pearson's chi-squared we estimate that $\phi = 16,232/599 = 20.32$, so the variance is about twenty times the mean. Allowing for over-dispersion the standard error of the coefficient of restaurant is $0.0312\sqrt{20.32} = 0.1406$. The revised test is $z = -0.3636/0.1406 = -2.586$, smaller but still significant at conventional levels.

- (c) Test the hypothesis that the negative binomial parameter called σ^2 in the notes and 'alpha' in the Stata output is zero. What would it mean if we were to accept the null hypothesis? What can you say about the sampling distribution of the likelihood ratio test criterion?

Twice the difference in negative log-likelihoods is $2(8111.52 - 1929.85) = 12,363.34$. A conservative test treats this statistic as a chi-squared with one d.f. and leads to overwhelming rejection of the null hypothesis. Accepting the null would mean that the distribution is Poisson, the special case of the negative binomial when $\sigma^2 = 0$. The asymptotic distribution of the test statistic is not chi-squared because the Poisson model falls in a boundary of the parameter space; a better approximation treats it as an average of zero and a chi-squared with one d.f. In this application the point is moot, however, because the test is overwhelmingly significant.

- (d) What's the expected number of cigarettes smoked per day for a non-white high-school graduate who is now 40 years old, makes \$22,026.47 and lives in a state where a pack of cigarettes costs \$5.46 if the state does not have restaurant restrictions? If it does?

We log income ($\log 22026.47 = 10$) and the price of cigarettes in cents ($\log 546 = 6.3$) and set education to 12 and age to 40. The linear predictor is 2.420 (see below) and exponentiating gives 11.24 cigs if there are no restaurant restrictions. The restrictions add -0.4704 to yield a linear predictor of 1.9493, and exponentiating gives 7.02 cigarettes. The basic calculation is $0.9635 - 0.2040(6.3) + 0.1025(10) - 0.0945(12) + 0.1360(40) - 0.0016(1600) = 2.449$ (rounded).

- (e) If someone was expected to smoke 8 cigarettes per day after taking all covariates into account, what would the variance be in the Poisson model? In the over-dispersed Poisson model? In the negative binomial model?

The Poisson variance is μ and we estimate it as 8. The over-dispersed Poisson variance is $\phi\mu$ and we estimate it as $20.32(8) = 162.6$. The negative binomial variance is $\mu(1 + \sigma^2\mu)$ and we estimate it as $8(1 + 7.369(8)) = 479.6$.

- (f) From the information available, which of the following models seems more appropriate for the data in terms of parsimony and goodness of fit: Poisson, negative binomial, or zero-inflated Poisson? How about the over-dispersed Poisson model?

The Poisson model clearly doesn't fit. The zero inflated Poisson and the negative binomial have the same number of parameters, so they tie in parsimony, but the negative binomial has a higher likelihood (-1929.85 vs. -2349.38), so it is clearly better. The over-dispersed Poisson model is obviously better than the Poisson; it has the same number of parameters as the other two, but we can't compare them in terms of likelihood because it doesn't have one.

[2] Mobility Limitations (35 pts)

A health survey conducted in Taiwan in 2006 asked a sample of 1279 older adults aged 53 to 98 whether they could jog a certain distance with no difficulty, with some or considerable difficulty, or were unable to do it. The outcomes are coded 0, 1 or 2 and the percents in each category are 64.8, 16.7 and 18.5%. Among the predictors of interest are age, years of education, and a dummy variable for females. For this analysis I centered age on 66 years and education on 6 years (complete primary), both values near the sample means. I also tried quadratic terms on age and education and, you will be happy to hear, discovered we really didn't need them.

- A multinomial logit model yields these results:

```

Multinomial logistic regression          Number of obs   =      1279
                                         LR chi2(6)      =      471.25
                                         Prob > chi2     =      0.0000
Log likelihood = -905.18072             Pseudo R2      =      0.2065

```

jog	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	

some/great						
agec	.0938395	.0091259	10.28	0.000	.075953 .1117261	
female	.4874911	.1798944	2.71	0.007	.1349046 .8400776	
educ	-.0648776	.0195267	-3.32	0.001	-.1031493 -.0266059	
_cons	-1.570264	.1283105	-12.24	0.000	-1.821748 -1.31878	

unable						
agec	.1541586	.0104555	14.74	0.000	.1336662 .174651	
female	.6257579	.1950452	3.21	0.001	.2434763 1.008039	
educ	-.1030498	.0212404	-4.85	0.000	-.1446802 -.0614194	
_cons	-2.037623	.1544388	-13.19	0.000	-2.340317 -1.734928	

(jog==no diff is the base outcome)

- A hierarchical logit model that first looks at whether respondents are able to jog and then whether they have (some or great) difficulty gives these results:

```

Logistic regression                    Number of obs   =      1279
                                         LR chi2(3)      =      306.63
                                         Prob > chi2     =      0.0000
Log likelihood = -459.7582             Pseudo R2      =      0.2501

```

able	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
agec	-.1218148	.0093771	-12.99	0.000	-.1401937 -.103436
female	-.4263757	.1806849	-2.36	0.018	-.7805115 -.0722399
educ	.080727	.0197673	4.08	0.000	.0419839 .1194702
_cons	2.21677	.1458199	15.20	0.000	1.930968 2.502571

```

Logistic regression                    Number of obs   =      1042
                                         LR chi2(3)      =      169.64
                                         Prob > chi2     =      0.0000
Log likelihood = -442.91415           Pseudo R2      =      0.1607

```

diff	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.0993353	.0096325	10.31	0.000	.0804559	.1182147
female	.539889	.1836257	2.94	0.003	.1799893	.8997887
educ	-.0637229	.0198953	-3.20	0.001	-.102717	-.0247287
_cons	-1.602115	.1307493	-12.25	0.000	-1.858379	-1.345851

- Finally, an ordered logit model yields these results:

Ordered logistic regression	Number of obs	=	1279
	LR chi2(3)	=	476.95
	Prob > chi2	=	0.0000
Log likelihood = -902.33287	Pseudo R2	=	0.2090

jog	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.1215326	.0072888	16.67	0.000	.1072468	.1358184
female	.5178293	.1411201	3.67	0.000	.2412391	.7944196
educ	-.0807107	.0153362	-5.26	0.000	-.1107691	-.0506522
/cut1	1.034822	.1046714			.8296701	1.239975
/cut2	2.261599	.1226008			2.021306	2.501892

- (a) Interpret the age coefficients in the multinomial logit model.

Comparing respondents with the same gender and education, the relative probability of having (some or great) difficulty jogging rather than none increases 9.8% per year of age, while the relative probability of being unable to jog rather than having no difficulty increases by 16.7% [$\exp(0.0938) - 1 = 0.0984$ and $\exp(0.1542) - 1 = 0.1667$].

- (b) What's the probability that a 66-year old man with 6 years of education would be unable to jog the required distance? How about a woman with the same age and education?

Note that age is centered on 66 and education on 6. For males all we need is the constants: $\exp(-2.0276) / (1 + \exp(-1.5703) + \exp(-2.0276)) = 0.974$ or 9.7%. For females we add the sex effects $\exp(-2.0276 + 0.6258) / (1 + \exp(-1.5703 + 0.4875) + \exp(-2.0276 + 0.6258)) = 0.154$ or 15.4%

- (c) Interpret the age coefficients in the hierarchical logit model noting carefully their signs.

Comparing respondents with the same gender and education, the odds of being able to jog *decrease* 11.5% per year of age [$\exp(-0.1218) - 1 = -0.1147$]. For those able to jog the odds of having (some or great) difficulty *increase* 10.4% per year of age [$\exp(0.0993) - 1 = 0.1044$].

- (d) Repeat the calculation of the probabilities in part (b) using the hierarchical logit model.

The probability of being unable to jog is obtained directly from the first equation as the complement. For males it is $1 - \text{logit}^{-1}(2.217) = 0.098$ or 9.8%. For females we add the sex coefficient to get $1 - \text{logit}^{-1}(2.217 - 0.426) = 0.143$ or 14.3%, both similar to part b.

- (e) Interpret the second cutpoint in the ordered logit model and the age coefficient in terms of odds and in terms of a latent variable representing mobility limitations.

The cut-points apply directly to 66-year old males with 6 years of education. For this group the (cumulative) odds of being able to jog, with or without difficulty, are $\exp(2.262) = 9.598$ or 9.6 to 1. In a standardized latent scale of mobility limitations this group falls above a z-score of 1.25 [$2.262/(\pi/\sqrt{3}) = 1.247$]. The age coefficient tells us that the odds of having difficulty or being unable to jog (relative to having no difficulty), as well as the odds of being unable to jog (relative to being able, with or without difficulty) increase 12.9% per year of age among males or females with the same education [$\exp(0.1215) - 1 = 0.129$]. In a latent scale of mobility limitations each year of age is associated with an increase of 0.07 standard deviations among males or females with the same education [$0.1215/(\pi/\sqrt{3}) = 0.070$].

- (f) Repeat the calculation of the probabilities in part (b) using the ordered logit model.

For males all we need is the second cutpoint, the probability is $1 - \text{logit}^{-1}(2.262) = 0.0944$ or 9.4%. For females we add the sex coefficient to get $1 - \text{logit}^{-1}(2.262 - 0.5178) = 0.1488$ or 14.9%, similar to parts b and d. *Note:* As usual, we change the sign of the regression coefficient (but not the cutpoint) to calculate cumulative probabilities. In this case I also subtract from one to get the probability of the last category, being unable to jog.

- (g) Which of the three models is more appropriate for the data at hand when you consider both parsimony and goodness of fit?

The ordered logit model has the highest log-likelihood (-902.33 , compared to -905.18 for the multinomial logit and -902.67 , obtained as $-459.76 - 442.91$, for the hierarchical logit), and it has the fewest parameters (only 5, compared to 8 for each of the other two models), so it is a clear winner. There is no need to compute the AIC because it wins on both parsimony and fit.

[3] Spells of Unemployment (35 pts)

Wichert and Wilke (2008) have data on a sample of 21,685 unemployment spells in West Germany starting in 1996 or 1997. A spell starts when an individual starts collecting unemployment benefits and ends with a transition into employment or with censoring on the last observed day of income transfers. The regressors of interest are age, gender (represented by an indicator for females), and the last daily wage before unemployment, in Euros (€). Time is measured in days but I divided by 365.25 to work with exposure in years.

- There are 18,615 “failures” (transitions into employment) in 26,386.73 person-years of observation. The last observed exit is just shy of six years.
- A piece-wise exponential model using splits at six months and 1, 2, 3 and 6 years has a log-likelihood of $-36,186.6$ and a model chi-squared of 2,355.5 on 4 d.f.
- Adding age, gender, and wage increased the model chi-squared to 3,688.8.
- The estimates of the baseline hazard show a much lower exit rate after six months with little variation after that, so I simplified the model by having just two duration categories, with a single dummy variable to identify segments starting at six months or later. Here’s the model:

Log likelihood = -35521.861	LR chi2(4) = 3684.98				
	Prob > chi2 = 0.0000				

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]

dur6mp	-.6561151	.0147816	-44.39	0.000	-.6850866 -.6271437
age	-.0243554	.0011119	-21.90	0.000	-.0265347 -.0221762
female	-.179553	.0158851	-11.30	0.000	-.2106873 -.1484188
wage	.0053406	.0002025	26.38	0.000	.0049437 .0057374
_cons	.6853334	.0413145	16.59	0.000	.6043586 .7663082

- (a) What's the average rate at which people leave unemployment? Assuming the rate is constant over time, what proportion would still be unemployed after one and two years? What's the median duration of unemployment, approximately?

The rate is $18,615/26,386.73 = 0.7055$ exits per person-year unemployed. At this rate we would expect $\exp(-0.7055) = 0.4939$ or 49.4% to be unemployed after a year and $\exp(-2(0.7055)) = 0.2439$ or 24.4% to be unemployed after two years. The median duration of unemployment is thus about a year. (The third quartile is about two years.)

- (b) Use the piecewise exponential model without covariates to verify that we have duration dependence. Can the decision to collapse the five duration categories into just two be justified using a likelihood ratio test? Explain.

There is significant duration dependence, the model chi-squared of 2,355.5 on 4 d.f. compares the piecewise and exponential models and is highly significant. We can justify the decision to collapse duration because the model with only two categories is nested in the model with five. We are not given the log-likelihoods without covariates, but after introducing age, gender and wages we have model chi-squareds of 3688.80 with five categories and 3684.98 with two, a difference of only 3.82 on three d.f., which is clearly not significant.

- (c) Describe the effects of age, gender and previous wages on the rate at which people leave unemployment. Because a difference of one euro is not very meaningful, report the effect of a difference of € 35, which is approximately the distance between the first and third quartiles.

The rate at which people leave unemployment *decreases* about 2.4% per year of age when we compare within the same gender, wages and duration of unemployment [$\exp(-0.0244) - 1 = -0.024$] or use the approx for small β]. The rate for women is 16.4% *less* than for men with the same age, wages and duration of unemployment [$\exp(-0.1796) - 1 = -0.1644$]. The rate for workers in the third quartile of the wage distribution is 20.6% *higher* than for workers in the first quartile with the same gender, age and time unemployed [$\exp(0.00534(35)) - 1 = 0.2055$].

- (d) Estimate the probability that a 35-year old male whose last wage was € 60 per day (both values close to the sample means) will still be unemployed after one and after two years. What's the probability of being unemployed after two years if the person has just completed one year unemployed?

The hazard is 1.166 in the first six months and 0.605 thereafter [log-hazard at 0-6 is $0.6853 - 0.0243(35) + 0.00534(60) = 0.1533$, $\exp(0.1533) = 1.166$ and $\exp(0.1533 - 0.6561) = 0.605$]. The

