

End Term

January 18, 2010

Please answer each question in a separate booklet. Take the 5% critical value of the normal distribution to be 2. For t , F and χ^2 tests report the statistic and the corresponding degrees of freedom. There's no need to calculate P -values.

[1] Adolescent Stress (30%)

Our first dataset has data on 651 adolescents who were asked about the number of stressful life events experienced in the last year. The predictors include a scale of family cohesion, a measure of self-esteem, past-year school grades, and a measure of school attachment (sattach), all coded from low to high.

- (a) The number of life events ranges from 0 to 10 and averages 2.6 per person. If the marginal distribution was Poisson what would the variance be? What would you make of the fact that the observed variance is 4? Under the same assumption, what proportion of the sample would experience no stressful events at all? The observed proportion is 12.9%. Comment.

Under Poisson the variance would be 2.6. The fact that it is 4 indicates over-dispersion. The probability of zero events in a Poisson distribution is $e^{-\mu}$, estimated as .074 or 7.4%. The fact that we observe 12.9% suggests excess zeroes (or zero inflation).

- (b) The enclosed Stata output shows the result of fitting a Poisson regression model with linear effects of the four predictors of interest. Note that all four predictors have significant net effects on the number of stressful events. Interpret the coefficients of family cohesion (which ranges from 18 to 75) and school attachment (range 8.2 to 36).

Family cohesion is associated with lower stress; each point in the cohesion scale is associated with one-percent fewer stressful events (or a 42% difference across the range). School attachment is also associated with reduced stress, each point corresponding also to about one-percent (1.18% to be precise) fewer events (or a 28% difference across the range).

- (c) The Pearson chi-squared statistic for the model of part (b) is 892.5 on 646 d.f. Assuming that the systematic part of the model is specified correctly, how would you interpret this result? Explain exactly how the conclusions of part (b) would be affected. (Hint: one conclusion would change.)

The ratio of the Pearson statistic to its degrees of freedom is $892.5/646 = 1.38$. Assuming no lack of fit, we attribute the fact that the ratio exceeds one to pure error (over-dispersion). This doesn't affect parameter estimates but the true standard errors are about $\sqrt{1.38} = 1.175$ times those reported. This affects the significance of the effect of grades, which would have a z -score of 1.69, below the usual 5% cutoff, but the other coefficients would still be significant.

- (d) Fitting a negative binomial regression model leads to a log-likelihood of -1273.81 and an estimate of the over-dispersion parameter (which Stata calls α and the notes call σ^2) of 0.153. Can you use this information to compare the Poisson and negative binomial models? Be specific.

The Poisson model is nested on the negative binomial, so we can compare the likelihoods of -1293.22 and -1273.81, which would give a chi-squared statistic of 38.82. This statistic does not have the usual chi-squared distribution with one d.f., but provides a conservative test and there is no question that the result is significant, so we reject the Poisson model. (Stata computes the p-value treating the statistic as a mixture of chi-squared statistics with one and zero d.f.)

- (e) What other model might one consider for this data? Describe the assumptions made by that model and how you would go about checking if it does a better job than the two alternatives considered so far.

The other model worth considering is a zero-inflated Poisson model, where some adolescents would not be exposed to stressful events, while the others would have a number of stressful events given by a Poisson distribution. The result of part a) suggests that this model might be better than the other two. We can check this by fitting the ZIP model and then (i) comparing the observed and predicted zeroes and (ii) comparing the maximized log-likelihoods while taking into account the number of parameters via AIC. (The Poisson is nested in ZIP, so a formal test is possible, with the usual precaution about models on a boundary of the parameter space.)

[2] Asset Allocation in Pension Plans (35%)

Papke collected data on the impact of allowing individuals to choose their own asset allocations in pension plans. The outcome of interest is whether the assets are “mostly bonds”, “mixed”, or “mostly stocks”. Choice is coded one if the person can choose how his or her pension fund is invested. Controls include age, education, gender, race, marital status, income (via a set of dummy variables), wealth, and whether the plan is profit sharing. The enclosed output shows some results from fitting multinomial, hierarchical and ordered logit models to the data. (I suppressed the coefficients of most control variables for simplicity. I also combined some of the income categories.)

- (a) Interpret the coefficients of choice in the multinomial logit model in terms of odds ratios or relative probabilities.

The odds (or relative probability) of having mostly bonds as opposed to a mixed portfolio are 53.9% lower, while the odds of having mostly stocks as opposed to a mixed portfolio are 8.1% higher, when people can choose asset allocation, everything else being equal ($e^{-0.7743} - 1 = -0.5390$ and $e^{0.0776} - 1 = 0.0807$.)

- (b) Evaluating the multinomial logit equations with every control variable set to the mean and choice set to zero we obtain log-odds of 0.3758 for bonds and -0.2082 for stocks. This translates

into a predicted probability for mostly stocks of 24.8%. What's the corresponding probability when choice is set to one?

The log-odds when choice is one are $0.3758 - 0.7743 = -0.3985$ for mostly bonds and $-0.2082 + 0.0776 = -0.1306$ for mostly stocks. The probability of a mostly stocks portfolio is then $e^{-0.1306} / (e^{-0.3989} + 1 + e^{-0.1306}) = 0.3443$ or 34.4%.

- (c) What would be the probability of mostly stocks according to this model if our average person had control of allocation but had to choose mostly stocks or mostly bonds, with mixed portfolios no longer an option?

According to the multinomial logit model the ratios of the expected utilities of stock and bonds would not change. The probability of mostly stocks for people with choice would be $e^{-0.1306} / (e^{-0.3989} + e^{-0.1306}) = 0.5666$ or 56.7%.

- (d) The hierarchical logit model looks first at whether people have a mixed portfolio (i.e. not mostly stocks or mostly bonds) and then, for those who do *not*, whether it is mostly stocks or mostly bonds. Interpret the coefficients of choice in this model.

According to the hierarchical logit model, the odds of having a mixed portfolio are 44.9% higher for people who have a choice ($e^{0.3711} = 1.4493$). The odds of having mostly bonds among those who do not have a mixed portfolio are 164% higher when people have a choice of asset allocation, everything else being equal ($e^{0.9718} = 2.6428$).

- (e) Evaluating the first and second stage equations with every control variable set to the mean and choice set to zero gives log-odds of -0.8461 for mixed portfolios and -0.6642 for mostly stocks among those without mixed portfolios. This translates into an unconditional probability of mostly stocks of 23.8%. Compute the corresponding probability when choice is set to one.

The log-odds when choice is one are $-0.8461 + 0.3711 = -0.4750$ for mixed portfolios and $-0.6642 + 0.9718 = 0.3076$ for mostly stocks given a portfolio that is *not* mixed. The probability of a mixed portfolio is then $\text{logit}^{-1}(-0.4750) = e^{-0.4750} / (1 + e^{-0.4750}) = 0.3834$ and the conditional probability of mostly stocks is $\text{logit}^{-1}(0.3076) = e^{0.3076} / (1 + e^{0.3076}) = 0.5763$. The unconditional probability of mostly stocks is then $(1 - 0.3834) \times 0.5763 = 0.3553$ or 35.5%.

- (f) Interpret the coefficient of choice in the ordered logit model in terms of odds and in terms of a latent variable ranging from preference for bonds to preference for stocks.

According to the ordered logit model, the odds of having *some* stocks (in a mixed or mostly stocks portfolio) rather than mostly bonds, as well as the odds of having *mostly* stocks rather than some bonds (in a mixed or mostly bonds portfolio) are 90% higher among people who have a choice of asset allocation, compared to those who do not, everything else being equal ($e^{0.6426} = 1.9015$). Alternatively, in a latent scale running from preference for bonds to preference for stocks, portfolios where people have a choice of asset allocation are 0.35 standard deviations further to the stock end of the scale ($0.6426 / (\pi / \sqrt{3}) = 0.3543$).

- (g) Which of these models best represents the effect of choice on asset allocation in terms of parsimony and goodness of fit?

The log-likelihoods are -191.625 for the multinomial logit, $-11.871 - 77.598 = -191.468$ for the hierarchical logit model, and -204.76 for the ordered logit model. The first two provide the best fit and are nearly indistinguishable. The ordered logit model uses 13 instead of 24 parameters so it is obviously more parsimonious. To compare non-nested models we use Akaike's information criterion, which gives 431.2, 430.9 and 435.5, respectively, so the ordered logit model is not acceptable.

[3] Physicians Who Smoke (35%)

Our last dataset for the day deals with coronary deaths among British physicians due to smoking. Imagine following physicians from age 35 (so all survival is conditional on having lived to age 35) up to age 84, although of course many are censored or die before that, and then tabulating events and exposure by smoking status and age in 10-year categories. The enclosed output shows a couple of survival models fitted using the standard Poisson trick with dummy variables representing the age categories, a dummy variable representing smoking status, and the log of exposure time as an offset.

- (a) Describe the shape of the baseline hazard in the piece-wise exponential proportional hazards model.

The hazard of coronary death increases monotonically (and rapidly) with age. Compared to ages 35-44, the risk is 4.4 times as high at 45-54, 13.8 times as high at 55-64, 28.5 times as high at 65-74 and 40.5 times as high at 75-84 ($\exp\{0, 1.4840, 2.6275, 3.3505, 3.7001\} = 4.411, 13.839, 28.517, 40.451$). Another way to express this is to take successive ratios and note that the risk increases fourfold in the first ten years, it triples in the next ten, doubles in the next ten, and increases just over forty percent in the last ten, so the proportionate increase is less in each decade (the ratios are 4.411, 3.138, 2.061, and 1.419)

- (b) Interpret the coefficient of smoking in the piece-wise exponential proportional hazards model and test its significance.

Physicians who smoke have a 42.4% higher risk of coronary death than non smoking physicians in the same age group. The relative risk is assumed to be the same in every age group. The Wald test is given in the output as $z = 3.3$, equivalent to $\chi^2 = 10.9$ on one d.f. and is significant at the one per thousand level (p-value=0.001).

- (c) Does the proportional hazards model fit the data? Justify your answer with a likelihood ratio test.

The proportional hazards model does not fit the data, the deviance or likelihood ratio chi-squared of 12.1 on 4.d. is significant at the five (but not the one) percent level (p-value = 0.016).

- (d) Describe the effect of smoking using the model that allows different effects at ages under 45, 45-64 and 65-84 (note the use of only 3 groups), and test whether the effect varies by age.

The effect of smoking is highest at 35-44, when the risk for smokers is 5.7 times that of non-smokers ($\exp\{1.7469\}=5.7366$), but declines to 1.7 times that of non-smokers at 45-64 ($\exp\{1.7469-1.2365\}=1.6659$) and 1.12 times that of non-smokers at ages 65-84 ($\exp\{1.7469-1.6293\}=1.1248$). Another way to look at this is to say that the *excess* risk of smokers is highest under age 45 but declines 70% at 45-64 ($\exp\{-1.237\}-1 = -0.710$) and 80% at 65-84 ($\exp\{-1.629\}-1 = -0.804$), always compared to under 45. Adding the interaction increases the model chi-squared by $931.9 - 922.93 = 9.06$ at the expense of two d.f., a significant change at the five (but not one) percent level (p -value, not required, is 0.011). So, the effect varies by age.

- (e) Does the last model fit the data? (I know I didn't give you the deviance, but you have enough information to compute it. Note that it is not zero.)

The deviance of this model can be obtained by subtracting the interaction chi-squared from the deviance of the additive model: $12.13 - 9.06 = 3.07$ on 2 d.f. (10 observations minus 8 parameters). This is clearly not significant, indicating that the model fits the data. (If we had used a separate interaction coefficient for each 10-year each group, instead of using 20-year groups as here, the model would be saturated, so we are really testing that we can use the simplified interaction term.)

- (f) What's the probability that a physician who doesn't smoke will survive from ages 35.0 to 85.0? (Check: nearly 70% do.) Compute the same probability for a physician who smokes.

The table below shows the log-hazard for a non-smoker, computed by adding the constant to each of the age coefficients. The hazard is obtained by exponentiation. The cumulative hazard is obtained multiplying by 10 and summing. The survival function is obtained by exponentiating the negative cumulative hazard. For smokers we repeat the process adding the appropriate smoking effect at each age. (The next column answers part g using the hazard for smokers under age 45 and the hazard for non-smokers thereafter.) The probability that a non-smoking physician will survive from 35 to 85 is 69.4%. The same probability for a smoker is 64.0%

Age	Non-smoker		Smoker		Quitter	
	$\log \lambda_0$	Survival	lrr	Survival	lrr	survival
35-44	-9.1479	0.9989	1.7469	0.9939	1.7469	0.9939
45-54	-6.5696	0.9850	0.5104	0.9710	0	0.9801
55-64	-5.4298	0.9428	0.5104	0.9026	0	0.9380
65-74	-4.3648	0.8302	0.1176	0.7823	0	0.8260
75-84	-4.0246	0.6943	0.1176	0.6398	0	0.6909

- (g) How much would it help to quit at age 45.0? To be precise, we want the probability of surviving from 35.0 to 85.0 for someone who plans to smoke from 35.0 to 45.0 and then quit, should he or she live that long.

Quitting at 45 helps, the probability is then 69.1%, very close to that of a non-smoker. (The future quitter gets away with the high excess risk under 45 because the baseline risk of coronary death at those ages is very low, one death per 10000 person-years.)

Stata Logs

[1] Adolescent Stress

```
. poisson stress cohes esteem grades sattach, nolog
```

```
Poisson regression                               Number of obs   =       651
                                                LR chi2(4)      =       82.45
                                                Prob > chi2     =       0.0000
Log likelihood = -1293.2219                    Pseudo R2      =       0.0309
```

stress	Coef.	Std. Err.	Z	P> z	[95% Conf. Interval]	
cohes	-.0096877	.0023806	-4.07	0.000	-.0143536	-.0050219
esteem	-.0171729	.0066112	-2.60	0.009	-.0301306	-.0042152
grades	-.0160849	.0080869	-1.99	0.047	-.0319349	-.000235
sattach	-.0119178	.0047713	-2.50	0.012	-.0212693	-.0025662
_cons	2.552202	.1941233	13.15	0.000	2.171727	2.932677

[2] Asset Allocation

```
. mlogit alloc choice `controls' // controls has the names of the control vars
```

```
Multinomial logistic regression                Number of obs   =       194
                                                LR chi2(22)    =       41.49
                                                Prob > chi2    =       0.0072
Log likelihood = -191.62503                    Pseudo R2      =       0.0977
```

alloc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mostly bonds						
choice	-.7742828	.4137461	-1.87	0.061	-1.58521	.0366446
age	.0749305	.0501705	1.49	0.135	-.0234019	.1732629
...						
prftshr	.3920406	.5296393	0.74	0.459	-.6460334	1.430115
_cons	-1.258517	3.396157	-0.37	0.711	-7.914864	5.397829
mostly sto~s						
choice	.0775942	.4268249	0.18	0.856	-.7589673	.9141557
age	-.0223388	.0538374	-0.41	0.678	-.1278581	.0831805
...						
prftshr	1.311086	.4996182	2.62	0.009	.3318518	2.290319
_cons	2.873469	3.62541	0.79	0.428	-4.232204	9.979142

(Outcome alloc==mixed is the comparison group)

. logit mixed choice `controls`

```

Logit estimates                                     Number of obs =      194
                                                    LR chi2(11)    =      28.17
                                                    Prob > chi2    =      0.0031
Log likelihood = -113.87056                       Pseudo R2     =      0.1101
    
```

mixed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
choice	.3711156	.360713	1.03	0.304	-.335869 1.0781
age	-.0308882	.0447603	-0.69	0.490	-.1186168 .0568405
...					
prftshr	-.8873557	.4575596	-1.94	0.052	-1.784156 .0094446
_cons	-1.307703	3.033961	-0.43	0.666	-7.254156 4.638751

. logit stock choice `controls` if !mixed

```

Logit estimates                                     Number of obs =      122
                                                    LR chi2(11)    =      13.64
                                                    Prob > chi2    =      0.2537
Log likelihood = -77.597631                       Pseudo R2     =      0.0808
    
```

stock	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
choice	.9718399	.4510347	2.15	0.031	.0878281 1.855852
age	-.0926428	.052339	-1.77	0.077	-.1952254 .0099397
...					
prftshr	1.080514	.4860645	2.22	0.026	.1278455 2.033183
_cons	3.988026	3.553836	1.12	0.262	-2.977364 10.95342

. ologit alloc choice `controls`

```

Ordered logit estimates                             Number of obs =      194
                                                    LR chi2(11)    =      15.22
                                                    Prob > chi2    =      0.1727
Log likelihood = -204.76137                       Pseudo R2     =      0.0358
    
```

alloc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
choice	.6426364	.2974384	2.16	0.031	.0596679 1.225605
age	-.0801431	.0385934	-2.08	0.038	-.1557848 -.0045015
...					
prftshr	.7931214	.3699815	2.14	0.032	.0679709 1.518272
_cut1	-4.139669	2.594853			(Ancillary parameters)
_cut2	-2.479187	2.582232			

[3] Physicians Who Smoke

```
. poisson deaths age45 age55 age65 age75 smokes, offset(logexpo)
```

```
Poisson regression                               Number of obs   =           10
                                                LR chi2(5)      =          922.93
                                                Prob > chi2     =           0.0000
Log likelihood = -33.600155                    Pseudo R2      =           0.9321
```

deaths	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age4554	1.484007	.1951034	7.61	0.000	1.101611 1.866402
age5564	2.627505	.1837273	14.30	0.000	2.267406 2.987604
age6574	3.350493	.1847992	18.13	0.000	2.988293 3.712692
age7584	3.700096	.1922195	19.25	0.000	3.323353 4.07684
smokes	.3545354	.1073741	3.30	0.001	.144086 .5649848
_cons	-7.919326	.1917618	-41.30	0.000	-8.295172 -7.543479
logexpo	(offset)				

```
. poisgof
```

```
Goodness-of-fit chi2 = 12.13244
Prob > chi2(4)      = 0.0164
```

```
. gen smokes4564 = smokes * (age45 + age55)
```

```
. gen smokes6584 = smokes * (age65 + age75)
```

```
. poisson deaths age45 age55 age65 age75 smokes*, offset(logexpo)
```

```
Poisson regression                               Number of obs   =           10
                                                LR chi2(7)      =          931.99
                                                Prob > chi2     =           0.0000
Log likelihood = -29.070159                    Pseudo R2      =           0.9413
```

deaths	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age4554	2.578303	.7280491	3.54	0.000	1.151353 4.005253
age5564	3.718156	.7258148	5.12	0.000	2.295585 5.140727
age6574	4.783111	.7206575	6.64	0.000	3.370648 6.195573
age7584	5.123332	.7216271	7.10	0.000	3.708969 6.537695
smokes	1.746873	.728869	2.40	0.017	.3183159 3.17543
smokes4564	-1.236502	.7480012	-1.65	0.098	-2.702558 .2295534
smokes6584	-1.629294	.7427882	-2.19	0.028	-3.085132 -.1734555
_cons	-9.147933	.7071068	-12.94	0.000	-10.53384 -7.762029
logexpo	(offset)				