End Term

December 2015 and January 2016

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, 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 do 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 table below shows the result of fitting a Poisson regression model with linear effects of the four predictors of interest yielding a log-likelihood of -1293.22. 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).

Variable	Coef	Std .Error	Z	P>|z|
Cohesion	-0.0097	0.0024	-4.07	0.000
Self-Esteem	-0.0172	.0066	-2.60	0.009
Grades	-0.0161	.0081	-1.99	0.047
Attachment	-0.0119	.0048	-2.50	0.012
Constant	2.5522	.1914		

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 we call σ^2 in the notes) 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] Self-Rated Health (35%)

Scott Lynch has data on self-rated health from the 1992 National Health and Nutrition Survey (NHANES). Health status is coded using 5 categories but for our purposes we will use just three: poor, fair, and good. The predictors are age in years (centered at 60), years of education (centered at 12) and indicators for females, blacks, and south. For simplicity we will consider an additive model only, although there is evidence that some interactions might be needed.

A *multinomial* logit model yields a log-likelihood of -3587.91 and the following estimates, using poor health as the baseline or reference category:

Variable	***Fair vs Poor***			**Good vs Poor**		
	Coef	Std. Error	Z	Coef	Std. Error	Z
Age – 60	-0.0225	0.0042	-5.31	-0.0439	0.0041	-10.84
Female	-0.0197	0.0985	-0.2	0.0148	0.0939	0.16
South	-0.2163	0.1052	-2.06	-0.3111	0.1007	-3.09
Black	-0.7293	0.1502	-4.86	-1.1565	0.1522	-7.60
Educ - 12	0.1191	0.0180	6.61	0.2159	0.0178	12.14
Constant	0.6813	0.0868	7.85	1.2192	0.0824	14.80

A *sequential* logit model that looks first at whether respondents have fair or good rather that poor health, and then whether they have good rather than fair health conditional on having either fair or good health, yields a combined log-likelihood of -3590.65 and the following estimates:

Variable	***Fair or Good vs Poor***			**Good vs Fair|Fair or Good**		
	Coef	Std. Error	Z	Coef	Std. Error	Z
Age – 60	-0.0348	0.0038	-9.26	-0.0211	0.0034	-6.17
Female	-0.0019	0.0876	-0.02	0.0330	0.0795	0.41
South	-0.2667	0.0930	-2.87	-0.0847	0.0895	-0.95
Black	-0.9567	0.1305	-7.33	-0.3851	0.1579	-2.44
Educ - 12	0.1718	0.0161	10.66	0.0911	0.0154	5.91
Constant	1.6911	0.0773	21.86	0.5347	0.0669	8.00

Finally an *ordered* logit model yields a log-likelihood of -3592.20 and the following estimates

Variable	**Coef**	**Std. Error**	**Z**
Age – 60	-0.0321	0.0028	-11.41
Female	0.0245	0.0653	0.37
South	-0.2107	0.0719	-2.93
Black	-0.8737	0.1150	-7.60
Educ - 12	0.1550	0.0124	12.47

Cutpoint	**Coef**
Poor|Fair	-1.6027
Fair|Good	-0.1116

(a) Interpret the coefficients of black in the multinomial logit model in terms of relative probabilities (a.k.a. relative odds).

The relative probability of being in fair rather than poor health is 52% lower, and the relative probability of good rather than poor health is 69% lower, for blacks than for non-blacks with the same age, gender, area of residence and education [exp(-0.7293, -1.1565) – 1 = (-0.518,-0.685)].

(b) What's the probability that a 60 year-old white male with 12 years of education who doesn't live in the south will report good health? How about a black male with the same age, education and region of residence?

The white target is the reference cell, so the linear predictors are the constants (0.6813,1.2192) so the probability is exp(1.2192)/(1+exp(0.6813)+exp(1.2192))=0.532 . For the black target we add the coefficients of black (-0.7293, -1.1565) to obtain (-0.048, 0.0627), so the probability is exp(0.0627)/(1 + exp(-0.048) + exp(0.0627)) = 0.353. The probability of reporting good health is 18 percentage points (or about 34%) lower for blacks than comparable whites.

(c) Interpret the coefficients of black in the sequential logit model.

The odds of reporting fair or good rather than poor health are 62% lower for blacks than comparable whites. Among those reporting fair or good health, the odds of good health are 32% lower for blacks than comparable whites, where comparable means with the same age, gender, education and residence. [Because exp{(-0.9567, -0.3851)} – 1 = (-0.6158, -0.3196).]

(d) Repeat the calculation of the probabilities of part (b) using the sequential logit model.

For white we need the constants c(1.6911, 0.5347); taking logit^{-1} we obtain step or conditional probabilities of 0.8444 and 0.6306. Multiplying these. the probability of good health is 0.532. For black we add the coefficients (-0.9567, -0.3851) to obtain (0.7344, 0.1496) which via logit^{-1} leads to step probabilities of 0.6758 and 0.5373; the product is 0.363. (Note similarity to part b.)

(e) Interpret the coefficient of black in the ordered logit model in terms of odds and in terms of a latent variable representing health status.

The coefficient is -0.8737. This means that the odds of good rather than fair or poor, as well as the odds of good or fair rather than poor health, are 58% lower for blacks than whites with the same age, gender, residence and education. [Because exp(-0.8737) - 1 = -0.5826.] It also means that the average health status of blacks is almost half a standard deviation lower than that of whites in a latent logistic scale, or 0.482 lower, to be precise. [Because $\frac{-0.8737}{\pi/\sqrt{3}} = -0.482$.]

(f) Repeat the calculation of the probabilities of part (b) using the ordered logit model. *Hint:* What does the second threshold parameter represent?

As before the white target is the reference cell, so the inverse logit of the second threshold is the cumulative probability of poor or fair health. The complement is $1 - \text{logit}^{-1}(-0.1116)$ = 0.5279. For black we add the coefficient with opposite sign to obtain a probability of $1 - \text{logit}^{-1}(-0.1116 + -0.8737)$ =0.3182. (Note the similarity to parts b and d.)

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

The multinomial logit has a thin edge over the sequential logit model, with a slightly higher log-likelihood. The ordered logit model fit has a somewhat worse fit with a lower log-likelihood by 8.58 points in the chi-square or deviance scale. But this model is also more parsimonious, with 7 rather than 12 parameters. The decision comes to a trade-off between parsimony and goodness of fit. AIC trades two chi-square points per parameter, so it would prefer the ordered logit model. (We don't know the sample size, but in all likelihood log(n) > 2, so BIC would concur.) We conclude that the ordered logit model represents the best compromise between parsimony and goodness of fit.

[3] Spells of Unemployment (35%)

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. The resulting model has a log-likelihood of −35521.86 and the following estimates

Variable	Coefficient	Std. Error	Z
Dur 6m+	-0.6561	0.0148	-44.39
Age	-0.0244	0.0011	-21.90
Female	-0.1796	0.0159	-11.30
Wage	0.0053	0.0002	26.38
Constant	0.6853	0.0413	16.59

(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} = .4939 or 49.4% to be unemployed after a year and exp{-2(0.7055)) = 0.244 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 ETA]. 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 conditional probability of remaining unemployed after two year if a 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 and exp{0.1533} = 1.166. Log hazard after 6 is 0.1533-0.6561= -0.5028 and exp(-0.5028) = 0.605]. The survival probabilities at one and two years are 0.4126 and 0.2253 [because exp{-.5(1.166) - 0.5(0.605)} = 0.4126 and exp{-0.5(1.166) - 1.5(0.605)} = 0.2253]. The conditional probability of remaining unemployed after two years given that the duration of unemployment is at least one year is 54.6% [calculated as 0.2252/0.4126 from first principles]

- At this point I decided to add an interaction between gender and duration by creating a dummy variable equal to the product of the indicators for Female and Dur 6m+, labeled Fem X Dur 6m+ below. This increased the log-likelihood to -35407.01 and produced the following estimates:

Variable	Coefficient	Std. Error	Z
Dur 6m+	-0.8248	0.0186	-44.38
Age	-0.0237	0.0011	-21.31
Female	-0.4389	0.0239	-18.32
Fem X Dur 6m+	0.4687	0.0312	15.04
Wage	0.0053	0.0002	26.21
Constant	0.7455	0.0414	18.01

(e) Verify that the effect of gender is in fact not proportional using a likelihood ratio test. Does the Wald test concur?

The difference in model chi-squareds (same as twice the difference in negative log-likelihoods) is 3914.68 - 3684.98 = 229.7 on one d.f. and is highly significant. The Wald test in the output is equivalent to a chi-squared of 15.04^2 =226.2, so it concurs.

(f) Note that the age and wage effects haven't really changed, but the other estimates have. Describe the duration and gender effects in light of the interaction as clearly and simply as you can.

In the first six months of unemployment the rate at which women find a job is 35.5% lower than the rate for men with the same age and previous wage [exp{0.4389} - 1= 0.355], but after that there are no appreciable gender differences [exp{-0.4389 + 0.4687} -1 = 0.03]. Looking at the duration estimates, the exit rate after six months unemployed declines 56.2% for men [exp{-0.8248} - 1 =0.5617] but only 30.0% for women with the same age and wages [exp{-0.8248 - 0.4687} - 1 = 0.2996].