Biostat 200C Final

Due June 16, 2023 @ 11:59PM

FirstName LastName (UID XXX-XXX-XXX)

Contents

This is an open book test. Helping or asking help from others is considered plagiarism.

Q1. (25 pts) Survival data analysis

Consider following survival times of 25 patients with no history of chronic diesease (`chr = 0`) and 25 patients with history of chronic disease (`chr = 1`).

1. Manually fill in the missing information in the following tables of ordered failure times for groups 1 (`chr = 0`) and 2 (`chr = 1`). Explain how survival probabilities (last column) are calculated.

Group 1 (`chr = 0`):

time	n.risk	n.event	survival
1.8	25	1	0.96
2.2	24	1	0.92
2.5	23	1	0.88
2.6	22	1	0.84
3.0	21	1	0.80
3.5	20	**???**	**???**
3.8	19	1	0.72
5.3	18	1	0.68
5.4	17	1	0.64
5.7	16	1	0.60
6.6	15	1	0.56
8.2	14	1	0.52
8.7	13	1	0.48
9.2	**???**	**???**	**???**
9.8	10	1	0.36
10.0	9	1	0.32
10.2	8	1	0.28
10.7	7	1	0.24
11.0	6	1	0.20
11.1	5	1	0.16
11.7	4	**???**	**???**

Group 2 (`chr = 1`):

time	n.risk	n.event	survival
1.4	25	1	0.96
1.6	24	1	0.92
1.8	23	1	0.88
2.4	22	1	0.84
2.8	21	1	0.80
2.9	20	1	0.76
3.1	19	1	0.72
3.5	18	1	0.68
3.6	17	1	0.64
3.9	???	???	???
4.1	???	???	???
4.2	???	???	???
4.7	13	1	0.48
4.9	12	1	0.44
5.2	11	1	0.40
5.8	10	1	0.36
5.9	9	1	0.32
6.5	8	1	0.28
7.8	7	1	0.24
8.3	6	1	0.20
8.4	5	1	0.16
8.8	4	1	0.12
9.1	???	???	0.08
9.9	???	???	0.04
11.4	1	1	0.00

2. Use R to display the Kaplan-Meier survival curves for groups 1 (`chr = 0`) and 2 (`chr = 1`).
3. Write down the log-likelihood of the parametric exponential (proportional hazard) model for survival times. Explain why this model can be fit as a generalized linear model with offset.
4. Fit the exponential (proportional hazard) model on the `chr` data using R. Interpret the coefficients.
5. Comment on the limitation of exponential model compared to other more flexible models such as Weibull.

Q2. (35 pts) Longitudinal data analysis

This question is adapted from Exercise 13.1 of ELMR (p294).

The `ohio` data concerns 536 children from Steubenville, Ohio and were taken as part of a study on the effects of air pollution. Children were in the study for 4 years from ages 7 to 10. The response was whether they wheezed (difficulty with breathing) or not. The variables are

- `resp`: an indicator of wheeze status (1 = yes, 0 = no)
- `id`: an identifier for the child
- `age`: 7 yrs = -2, 8 yrs = -1, 9 yrs = 0, 10 yrs = 1
- `smoke`: an indicator of maternal smoking status at the first year of the study (1 = smoker, 0 = non-smoker).

1. Construct a table that shows proportion of children who wheeze for 0, 1, 2, 3 or 4 years broken down by maternal smoking status.
2. Make a plot which shows how the proportion of children wheezing changes by age with a separate line for smoking and nonsmoking mothers.
3. Group the data by child to count the total (out of four) years of wheezing. Fit a binomial GLM to this response to check for a maternal smoking effect. Does this prove there is a smoking effect or could there be another plausible explanation? Discuss the potential issue of using GLM here.
4. Fit a model for each individual response using a GLMM fit using penalized quasi-likelihood. Interpret the coefficients of age and maternal smoking. How do the odds of wheezing change numerically over time? Explain what you observe in the output AIC, BIC, and logLik.
5. Now fit the same model but using adaptive Gaussian-Hermit quadrature (with 25 knots). Compare to the previous model fit. Interpret the fixed effect coefficients. Interpret the variance component estimates. What is the odds ratio for wheezing comparing a smoking mother (`smoke = 1`) vs a non-smoking mother (`smoke = 0`) and its 95% confidence interval.
6. Fit the model using GEE. Use an autoregressive rather than exchangeable error structure. Explain the underlying assumptions of the autoregressive correlation structure. Does the GLMM approach in 5 assumes the same correlation structure within each cluster/individual? Compare the results to the previous model fits. In your model, what indicates that a child who already wheezes is likely to continue to wheeze? What happens if we misspecified the correlation structure?
7. What is your overall conclusion regarding the effect of age and maternal smoking? Can we trust the GLM result or are the GLMM models preferable?

Q3. (40 pts) LMM and GAMM

This question is adapted from Exercise 11.2 of ELMR (p251). Read the documentation of the dataset `hprice` in Faraway package before working on this problem.

1. Make a plot of the data on a single panel to show how housing prices increase by year. Describe what can be seen in the plot.
2. Fit a linear model with the (log) house price as the response and all other variables (except msa) as fixed effect predictors. Which terms are statistically significant? Discuss the coefficient for time.
3. Make a plot that shows how per-capita income changes over time. What is the nature of the increase? Make a similar plot to show how income growth changes over time. Comment on the plot.
4. Create a new variable that is the per-capita income for the first time period for each MSA. Refit the same linear model but now using the initial income and not the income as it changes over time. Compare the two models.
5. Fit a mixed effects model that has a random intercept for each MSA. Why might this be reasonable? The rest of the model should have the same structure as in the previous question. Make a numerical interpretation of the coefficient of time in your model. Explain the difference between REML and MLE methods.
6. Fit a model that omits the adjacent to water and rent control predictors. Test whether this reduction in the model can be supported.
7. It is possible that the increase in prices may not be linear in year. Fit an additive mixed model where smooth is added to year. Make a plot to show how prices have increased over time.
8. Interpret the coefficients in the previous model for the initial annual income, growth and regulation predictors.

Optional Extra Credit Problem*

> This problem is meant to offer another chance to demonstrate understanding of some of the material on the mid-term. If you choose to do this problem and your score is higher than your mid-term grade, then your mid-term grade will be reweighted to be `New Midterm Grade = .8*Old Midterm Grade + .2*Extra Credit Problem`

The following table shows numbers of beetles dead after five hours exposure to gaseous carbon disulphide at various concentrations.

```
(beetle <- tibble(dose = c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839),
                  beetles = c(59, 60, 62, 56, 63, 59, 62, 60),
                  killed = c(6, 13, 18, 28, 52, 53, 61, 60)))
```

```
## # A tibble: 8 x 3
##    dose beetles killed
##   <dbl>   <dbl>  <dbl>
## 1  1.69      59      6
## 2  1.72      60     13
## 3  1.76      62     18
## 4  1.78      56     28
## 5  1.81      63     52
## 6  1.84      59     53
## 7  1.86      62     61
## 8  1.88      60     60
```

1. Let x_i be `dose`, n_i be the number of beetles, and y_i be the number of killed. Plot the proportions $p_i = y_i/n_i$ plotted against dose x_i.
2. We fit a logistic model to understand the relationship between dose and the probably of being killed. Write out the logistic model and associated log-likelihood function.
3. Derive the scores, $\mathbf{U}$, with respect to parameters in the above logistic model. (Hint there are two parameters)
4. Derive the information matrix, $\mathcal{I}$ (Hint, a 2×2 matrix)
5. Maximum likelihood estimates are obtained by solving the iterative equation

$$\mathcal{I}^{(m-1)}\mathbf{b}^{(m)} = \mathcal{I}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

where $\mathbf{b}$ is the vector of estimates. Starting with $\mathbf{b}^{(0)} = 0$, implement this algorithm to show successive iterations are

Iterations	β_1	β_2	log-likelihood
0	0	0	-333.404
1	-37.856	21.337	-200.010
2	-53.853	30.384	-187.274
3			
4			
5			
6	-60.717	34.270	-186.235

- If after 6 steps, the model converged. For this final model, calculate the deviance. What is the distribution the deviance has?
- Does the model fit the data well? justify your answer.