BIOSTAT/STAT 570: Final

To be submitted to the course canvas site by 11:59pm Wednesday 15th December, 2021.

1. The data in Table 1 contains information on 4334 individuals who were surveyed in different regions of the country and asked their religious beliefs.

	Religious Beliefs			
Region	Fundamentalist	Moderate	Liberal	Total
Northeast	92	352	234	678
Midwest	274	399	326	999
South	739	536	412	1687
West/Mountain	192	423	355	970
All	1297	1710	1327	4334

Table 1: Data on region of residence and religious beliefs in the United States. From the 2006 General Social Survey.

We let i index the rows of Table 1 and j index the columns, so that the random variables Y_{ij} represent the number of individuals in row i, $i = 1, \dots, 4$, and in category j, with $j = 0, 1, 2$, corresponding to Fundamentalist/Moderate/Liberal. Let

$$p_{ij} = \Pr(\text{ Response } = j | \text{ Region } = i), \qquad j = 0, 1, 2,$$

for $i = 1, 2, 3, 4$.

The aim of the analysis is to understand how the different levels of religious beliefs are associated with region.

(a) **5 marks** Provide a single plot that shows the association between p_{ij} and region i, and comment on the plot.

(b) **10 marks** In our first analysis we collapse columns 2 and 3 and let $Z_i = Y_{i0}$ and

$$q_i = \Pr(\text{ Response = 0 } | \text{ Region } = i).$$

Suppose $Z_i | q_i \sim_{ind}$ Binomial(N_i, q_i), for $i = 1, \dots, n$, and consider the logistic regression model,

$$\log \left(\frac{q_i}{1 - q_i} \right) = \gamma_i \qquad (1)$$

and write $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3, \gamma_4]^\mathsf{T}$. Write down the likelihood $L(\boldsymbol{\gamma})$ for the sample z_i, $i = 1, 2, 3, 4$.

(c) **10 marks** Fit the model described in the previous part, and give asymptotic 95% confidence intervals for the odds ratios. Carefully interpret the γ_i parameters, for $i = 1, 2, 3, 4$.

(d) **5 marks** Is the fit of model (1) a significant improvement over the model in which q_i does not depend on region?

(e) **10 marks** We will now consider analyses that do not coarsen the data, via the model

$$\boldsymbol{Y}_i|\boldsymbol{p}_i \sim_{ind} \text{Multinomial}(N_i, \boldsymbol{p}_i),$$

where $\boldsymbol{p}_i = [p_{i0}, p_{i1}, p_{i2}]^\mathsf{T}$. For ordinal data, such as those in Table 1, we can consider *proportional odds models* with

$$\pi_{ij} = \Pr(\text{ Response } \leq j | \text{ Region } = i),$$

and

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_j - \beta_i,$$

for $i = 1, 2, 3, 4$ and $j = 0, 1$. Let $\boldsymbol{\alpha} = [\alpha_0, \alpha_1]^\mathsf{T}$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4]^\mathsf{T}$ where for identifiability we take $\beta_1 = 0$. Write down, in as simplified a form as possible, the log-likelihood $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for the sample y_{ij}, $i = 1, 2, 3, 4$, $j = 0, 1, 2$.

(f) **15 marks** Fit the proportional odds models:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_j \tag{2}$$

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_j^\star - \beta_i^\star \tag{3}$$

using the `polr` function in the `MASS` library. Compare these models using a likelihood ratio test, and summarize the association between religious beliefs and region, using your favored model.

(g) **10 marks** Provide a plot of fitted probabilities under model (3), as a function of region. Examine the fit (using any additional plots/analyses if you think they are useful) – does your preferred model provide an adequate fit?

2. Consider multinomial data with J categories and n different conditions $\boldsymbol{Y}_i = [Y_{i0}, \ldots, Y_{i,J-1}]$, $i = 1, \ldots, n$, with

$$\boldsymbol{Y}_i | \boldsymbol{p}_i \sim \text{Multinomial}_J(N_i, \boldsymbol{p}_i), \tag{4}$$

with $\boldsymbol{p}_i = [p_{i0}, \ldots, p_{i,J-1}]$.

(a) **6 marks** Suppose the response categories are nominal, i.e., have no ordering. In this case, a common approach is to fit the *generalized logit model*. With a single covariate x, this model has the form,

$$\log\left(\frac{p_{ij}}{p_{i,J-1}}\right) = \alpha_j + x_i \beta_j, \tag{5}$$

for $j = 1, \ldots, J-2$, so that we are modeling the odds, relative to the last category. For identifiability, we take $\alpha_{J-1} = \beta_{J-1} = 0$.
Show that this form is equivalent to,

$$p_{ij} = \frac{\exp(\alpha_j + x_i \beta_j)}{\sum_{j'=0}^{J-1} \exp(\alpha_{j'} + x_i \beta_{j'})}, \qquad j = 0, \ldots, J-1,$$

for $i = 1, \ldots, n$.

(b) **10 marks** Table 2 contains data from 3165 women who took part in the Demographic Health Survey (DHS) in El Salvador in 1985. The data are grouped in 5-year intervals and give the current use of contraception, classified as sterilization, other methods and no method.

Table 2: Data from El Salvador DHS in 1985.

	Contraceptive Method			
Age	Ster.	Other	None	Total
15–19	4	61	232	296
20–24	80	137	400	617
25–29	216	131	301	648
30–34	268	76	203	547
35–39	197	50	188	435
40–44	150	24	164	338
45–49	91	10	183	284
All	1005	489	1671	3165

For these data we have $J = 3$ levels (with $j = 0$ and $j = 1$ corresponding to sterilization and other, respectively) and a single covariate x, which we take as the age category mid-point. With respect to the model (5), give interpretations of $\exp(\alpha_j)$ and $\exp(\beta_j)$, $j = 0, 1$, with respect to the El Salvador data context.

(c) **10 marks** For these data provide a plot of the empirical generalized logits (as described by (5)) versus x.

(d) **20 marks** Using the `multinom` function in the `nnet` package fit the models

$$\begin{aligned} \log\left(\frac{p_{ij}}{p_{iJ}}\right) &= \alpha_j + x_i\beta_j \\ \log\left(\frac{p_{ij}}{p_{iJ}}\right) &= \alpha_j^\star + x_i\beta_j^\star + x_i^2\gamma_j^\star \end{aligned}$$

for $j = 0, 1$, where x_i is the mid-point of the woman's age band.

(e) **6 marks** For your preferred model, again plot the empirical logits, and add the fitted lines (which will either be linear or quadratic). Which model is preferred?

(f) **8 marks** Consider again the model given by (4) and (6) and suppose the probabilities are modeled as

$$p_{ij} = \frac{g_{ij}(\boldsymbol{\beta})}{G_i(\boldsymbol{\beta})}, \tag{6}$$

where $G_i(\boldsymbol{\beta}) = \sum_{j'=0}^{J-1} g_{ij'}(\boldsymbol{\beta})$, with unknown parameters $\boldsymbol{\beta}$. The MLE's of $\boldsymbol{\beta}$, are found by maximizing the likelihood:

$$L_M(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=0}^{J-1}\left[\frac{g_{ij}(\boldsymbol{\beta})}{G_i(\boldsymbol{\beta})}\right]^{y_{ij}}.$$

Now consider the model

$$Y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij}),$$

with

$$\mu_{ij} = \exp(\phi_i) \times g_{ij}(\boldsymbol{\beta}),$$

for $i = 1, \ldots, n$, $j = 0, \ldots, J-1$. Write down the likelihood function $L_P(\boldsymbol{\phi}, \boldsymbol{\beta})$, where $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]$.

(g) **6 marks** Find the MLEs $\widehat{\phi}_i = \widehat{\phi}_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$.

(h) **8 marks** Plug the MLEs $\widehat{\phi}_i$ into $L_P(\boldsymbol{\phi}, \boldsymbol{\beta})$, and show that

$$L_P(\widehat{\boldsymbol{\phi}}(\boldsymbol{\beta}), \boldsymbol{\beta}) \propto L_M(\boldsymbol{\beta}).$$

Hence, explain how one might fit the model given by (4) and (6) using a Poisson sampling model.

(i) **Bonus 15 marks** Using the `multinom` function, fit the multinomial model with a factor for each age group. Fit the equivalent Poisson log-linear model, and show that the same parameter estimates are recovered.