STA 325: Homework 5

DUE: 11:59pm, December 8 (on Sakai)
COVERAGE: ISL Chapter 9, Multicategory Modeling

1. **[30 points]** ISL Chapter 9, Question 8.

2. **[20 points]** Let's take a closer look at the baseline-logit model, which extends logistic regression for nominal categorical responses. Suppose we have J categories which are unordered. Let $Y(x) \in \{1, \cdots, J\}$ be the random variable for the categorical response at a single predictor x, and let $\pi_j(x) = \mathbb{P}[Y(x) = j]$, $j = 1, \cdots, J$. Using the first category as a baseline, the baseline-logit model assumes:

$$\log\left(\frac{\pi_j(x)}{\pi_1(x)}\right) = \alpha_j + \beta_j x, \quad j = 2, \cdots, J. \tag{1}$$

   (a) **[6 points]** Suppose the slope coefficients $\beta_2 = -0.1$ and $\beta_3 = 0.5$. Interpret these two coefficients for the baseline-logit model. What can you say about the log-odds $\log(\pi_2(x)/\pi_3(x))$ as x changes? (*Hint*: see Equation (6.2) in notes.)

   (b) **[6 points]** Show that the probabilities from the baseline-logit model (1) follow:

$$\pi_j(x) = \frac{\exp\{\alpha_j + \beta_j x\}}{\sum_{k=1}^J \exp\{\alpha_k + \beta_k x\}}, \quad j = 1, \cdots, J,$$

   where $\alpha_1 = \beta_1 = 0$. (*Hint*: Solve (1) in terms of $\pi_j(x)$, then sum together the equations for $j = 2, \cdots, J$. Using the fact that $1 - \sum_{j=2}^J \pi_j(x) = \pi_1(x)$, solve the resulting equation in terms of $\pi_1(x)$, then use this to get the expression for $\pi_j(x)$.)

   (c) **[8 points]** Suppose we are interested in how a student's score on a writing test is associated with the type of program he / she is in ("General", "Academic" or "Vocational"). The first quantity is measured by the continuous variable `write`, and the second is measured by the nominal categorical variable `prog` with levels 1 (baseline), 2 and 3. R gives the following output on the model fit `model <- multinom(prog ~ write)`:

```
> summary(model)
Call:
multinom(formula = prog ~ write, data = mdata)

Coefficients:
  (Intercept)      write
2   -2.712081  0.0660027
3    2.646492 -0.0517980

Std. Errors:
  (Intercept)      write
2    1.132812 0.02100153
3    1.127455 0.02251449

Residual Deviance: 371.0217
AIC: 379.0217
```

   Using this output, plot out the fitted probabilities $\pi_1($`write`$)$, $\pi_2($`write`$)$ and $\pi_3($`write`$)$, as `write` changes from a minimum score of 31 to a maximum score of 67. Interpret your results in terms of the problem.

3. **[24 points]** Let's take a closer look at the cumulative-logit model, which extends logistic regression for ordinal categorical responses. Suppose we have J categories which are ordered. Let $Y(x) \in \{1, \cdots, J\}$ be the random variable for the categorical response at a single predictor x, and let $\bar{\pi}_j(x) = \mathbb{P}[Y(x) \leq j]$, $j = 1, \cdots, J$. The proportional-odds cumulative-logit model (as implemented in the R function `polr`) assumes:

$$\text{logit}\{\bar{\pi}_j(x)\} = \alpha_j - \beta x, \quad j = 1, \cdots, J-1. \tag{2}$$

(a) **[6 points]** Give an expression for the cumulative probabilities $\bar{\pi}_j(x)$, $j = 1, \cdots, J-1$, and also for $\bar{\pi}_J(x)$. Using this, give an expression for the class probabilities $\pi_j(x)$, $j = 1, \cdots, J$.

(b) **[10 points]** Suppose we have $J = 4$ ordered classes, with $\alpha_1 = -1$, $\alpha_2 = 0$, $\alpha_3 = 1$ and $\beta = -0.5$. Plot the cumulative probabilities $\bar{\pi}_j(x)$, $j = 1, \cdots, 3$ and the class probabilities $\pi_j(x)$, $j = 1, \cdots, 4$ as a function of x. Interpret the slope parameter $\beta = -0.5$. How do the class probabilities change as x increases?

(c) **[8 points]** Suppose we are interested in how a college junior's GPA influences whether or not he / she applies to graduate school. The first quantity is measured by the continuous variable `gpa`, and the second is measured by the ordinal categorical variable `apply` with levels 1 (unlikely), 2 (somewhat likely) and 3 (very likely). R gives the following output on the model fit `model <- polr(apply ~ gpa)`:

```
> summary(model)
Call:
polr(formula = apply ~ gpa, data = dat, Hess = TRUE)

Coefficients:
     Value Std. Error t value
gpa 0.7249     0.2493   2.908

Intercepts:
                            Value  Std. Error t value
unlikely|somewhat likely    2.3748 0.7570     3.1372
somewhat likely|very likely 4.3998 0.7815     5.6301

Residual Deviance: 732.60
AIC: 738.60
```

Using this output, plot out the fitted probabilities $\pi_1(\texttt{gpa})$, $\pi_2(\texttt{gpa})$ and $\pi_3(\texttt{gpa})$, as `gpa` changes 2.0 to 4.0. Interpret your results in terms of the problem.

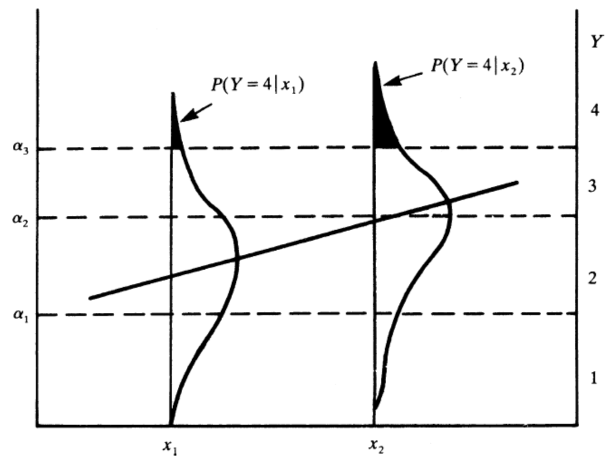
(d) **[BONUS 4 points]** Suppose the underlying generating mechanism for the categorical response $Y(x)$ follows the two-step procedure. First, for a fixed value of x, simulate the latent (i.e., unobserved) random variable $Z(x)$ from the normal distribution:

$$Z(x) = \mathcal{N}(\beta x, 1).$$

Next, let $Y(x)$ be the following discretization of $Z(x)$:

$$Y(x) = \begin{cases} 1, & \text{if } Z(x) \in (-\infty, \alpha_1] \\ 2, & \text{if } Z(x) \in (\alpha_1, \alpha_2] \\ 3, & \text{if } Z(x) \in (\alpha_2, \alpha_3] \\ \vdots & \\ J, & \text{if } Z(x) \in (\alpha_{J-1}, +\infty) \end{cases}.$$

Another way to view this is that $Y(x)$ is *binned* into the predetermined intervals $(-\infty, \alpha_1]$, $(\alpha_1, \alpha_2]$, ... , $(\alpha_{J-1}, +\infty)$; see below:



Under this generative model, show that the cumulative probabilities $\bar{\pi}_j(x)$ follow:

$$G\{\bar{\pi}_j(x)\} = \alpha_j - \beta x, \quad j = 1, \cdots, J-1,$$

for some function G (specify what G is). Compare and contrast this new model with the model in (2). What needs to be changed in the generating mechanism to yield the cumulative logit model (2)?