<table>
<thead>
<tr>
<th>Question</th>
<th>Possible Points</th>
<th>Actual Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>
1. The following data are counts of road accidents in Maine, classified according to location and gender of the patients, and whether the person was wearing a seat belt. The response categories are (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services, (4) injured, hospitalized and survived, (5) injured and not survived.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Location</th>
<th>Seat Belt</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Urban</td>
<td>No</td>
<td>7287</td>
<td>175</td>
<td>720</td>
<td>91</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>11587</td>
<td>126</td>
<td>577</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>Rural</td>
<td>No</td>
<td>3246</td>
<td>73</td>
<td>710</td>
<td>159</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>6134</td>
<td>94</td>
<td>564</td>
<td>82</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>Urban</td>
<td>No</td>
<td>10381</td>
<td>136</td>
<td>566</td>
<td>96</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>10969</td>
<td>83</td>
<td>259</td>
<td>37</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>No</td>
<td>6123</td>
<td>141</td>
<td>710</td>
<td>188</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>6693</td>
<td>74</td>
<td>353</td>
<td>74</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

A proportional odds model was fit to the response, using the remaining categories as predictors. Parameter estimates and standard errors of the fit are given below. All parameters have the correct signs (i.e. you do not need to take the negative value of parameter estimates, as is usually done with the \texttt{polr} output).

<table>
<thead>
<tr>
<th>Coefficients:</th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept):1</td>
<td>3.30742</td>
<td>0.035102</td>
<td>94.2233</td>
</tr>
<tr>
<td>(Intercept):2</td>
<td>3.48185</td>
<td>0.035562</td>
<td>97.9103</td>
</tr>
<tr>
<td>(Intercept):3</td>
<td>5.34938</td>
<td>0.046950</td>
<td>113.9385</td>
</tr>
<tr>
<td>(Intercept):4</td>
<td>7.25633</td>
<td>0.091428</td>
<td>79.3664</td>
</tr>
<tr>
<td>genderfemale</td>
<td>-0.54625</td>
<td>0.027211</td>
<td>-20.0748</td>
</tr>
<tr>
<td>seatbeltno</td>
<td>-0.76016</td>
<td>0.039382</td>
<td>-19.3021</td>
</tr>
<tr>
<td>locationrural</td>
<td>-0.69885</td>
<td>0.042388</td>
<td>-16.4867</td>
</tr>
<tr>
<td>seatbeltno:locationrural</td>
<td>-0.12442</td>
<td>0.054765</td>
<td>-2.2718</td>
</tr>
</tbody>
</table>

(a) \textbf{(6 pts)} State the model and the distributional assumptions.

\textbf{Answer:} $Y|\mathbf{X} \sim \text{Multinomial}(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$. If we define $P_j = \Pr(Y \leq j | \mathbf{X})$, then the model is

\[
\log \left( \frac{P_j}{1 - P_j} \right) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{23} X_2 X_3, \quad j = 1, \ldots, 4
\]

where $X_1 = I(\text{gender=female})$, $X_2 = I(\text{seatbelt=no})$ and $X_3 = I(\text{location=rural})$. \hfill \(3\)
(b) (6 pts) For males in urban areas wearing seat belts, calculate the estimated probabilities of the first two response categories.

**Answer:**

The cumulative probabilities are

\[ P_j = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \]

and the values for all the categories are

\[
\begin{pmatrix}
\exp(3.30742) & \exp(3.48185) & \exp(5.34938) & \exp(7.25633)
\end{pmatrix} \left( \begin{pmatrix}
1+\exp(3.30742) & 1+\exp(3.48185) & 1+\exp(5.34938) & 1+\exp(7.25633)
\end{pmatrix}^{-1}
\right)^{-1}
\]

The response probabilities are

\[ \pi_1 = P_1, \text{ and } \pi_j = P_j - P_{j-1} \text{ for } j = 2, \ldots, 5 \]

therefore the values for all the categories are (0.9647, 0.0055, 0.0251, 0.0040, 0.0007)

(c) (6 pts) Give a point estimate of the cumulative odds ratio of the gender, given seat belt use and location. Interpret this estimate.

**Answer:**

The odds ratio is defined as

\[ OR = \frac{\text{odds } P\{Y \leq j | \text{female}\}}{\text{odds } P\{Y \leq j | \text{male}\}} \]

The point estimate is \( \hat{OR} = \exp(\beta_1) = \exp(-0.54625) = 0.579 \). Given any status of seat belt use and location, the estimated odds of injury below any fixed level for a female are 0.579 times the estimated odds for a male.
(d) (6 pts) Find and interpret the estimated cumulative odds ratio between response and seat belt use, given that the accidents occurred in a rural location. Why is this estimate different from the estimate for the accidents in urban locations?

**Answer:**

The odds ratio is defined as

$$OR = \frac{\text{odds} P\{Y \leq j | \text{seatbelt = no, rural}\}}{\text{odds} P\{Y \leq j | \text{seatbelt = yes, rural}\}}$$

The point estimate is $\hat{OR} = \exp(\beta_2 + \beta_{23}) = \exp(-0.76016 - 0.12442) = 0.41$ in rural locations and $\exp(-0.76016) = 0.47$ in urban locations. The interaction effect $-0.1244$ represents the difference between the log odds ratios in rural and urban locations.
2. Consider an $I \times J \times 2$ table with variables $X$, $W$ and $Y$ respectively. Assume that the counts in the table are consistent with the Poisson sampling.

(a) (6 pts) Specify a log-linear model that assumes an association between each pair of the variables for each level of the third variable, i.e. $(XY, WY, XW)$. State the model, the assumptions, the constraints and the model degrees of freedom.

**Answer:**

We assume that the count in each cell $Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$, where $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, 2$ and

$$
\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk},
$$

$$
\mu = \log E\{Y_{111}\},
$$

$$
\alpha_1 = \beta_1 = \gamma_k = \alpha\gamma_{1k} = \alpha\gamma_{i1} = \beta\gamma_{1k} = \beta\gamma_{j1} = 0
$$

The model df = $1 + (I-1) + (J-1) + (2-1) + (I-1)(J-1) + (I-1)(2-1) + (J-1)(2-1) = IJ + I + J - 1$

Equivalently, the model df can be obtained as $IJ2 - (I - 1)(J - 1)(2 - 1)$

(b) (6 pts) Specify the equivalent logistic regression model with $Y$ as the response. State the model, the assumptions, the constraints and the model degrees of freedom.

**Answer:**

The equivalent logistic regression model is

$$
Y_{2ij} \sim \text{Binomial}(n_{2ij}, \pi_{2ij}), \text{ where }
$$

$$
\log \left( \frac{\pi_{2ij}}{1 - \pi_{2ij}} \right) = \mu + \alpha_i + \beta_j; \quad \mu = \log \left( \frac{\pi_{211}}{1 - \pi_{211}} \right) \text{ and } \alpha_1 = \beta_1 = 0
$$

The model df = $1 + (I-1) + (J-1) = I + J - 1$. Since the model conditions on $n_{ij}$, this accounts for the additional $IJ$ degrees of freedom. Therefore the two models are equivalent.
(c) (6 pts) Denote the fitted cell counts according to the log-linear model \( \hat{\lambda}_{ijk} \) and the fitted probabilities according to the logistic regression \( \hat{\pi}_{1|ij} \) and \( \hat{\pi}_{2|ij} \). Write \( \hat{\pi}_{2|ij} \) in terms of \( \hat{\lambda}_{ijk} \).

**Answer:**

\[
\hat{\pi}_{2|ij} = \frac{\hat{\lambda}_{ij2}}{\hat{\lambda}_{ij1} + \hat{\lambda}_{ij2}}
\]

(d) (6 pts) Show that according to this model (in either formulation (a) or (b), since these are equivalent) the odds ratios of levels of \( Y \) given different levels of \( X \) are independent of \( W \).

**Answer:** E.g., using the formulation in (b):

\[
\log \text{OR} = \log \frac{\text{odds}(Y = 2 \mid X = i', W = j)}{\text{odds}(Y = 2 \mid X = i, W = j)} = \log \left( \frac{\pi_{2|i'j}/(1 - \pi_{2|i'j})}{\pi_{2|ij}/(1 - \pi_{2|ij})} \right) = \alpha_{i'} - \alpha_i
\]
3. The geometric probability distribution can be interpreted as the distribution of the number of \textit{iid} Bernoulli trials observed until a first success. Its probability mass function is

\[ f(y \mid \pi) = \pi^y (1 - \pi), \quad y = 0, 1, 2, \ldots \]

(a) (6 pts) Show that this is a member of the exponential family of distributions. According to the notation used in class, state \( \phi, a(\phi), \theta, b(\theta) \) and \( c(y, \phi) \).

\textbf{Answer:}

\[ f(y \mid \pi) = \exp(y \log \pi + \log(1 - \pi)) \]

Therefore

\[ \theta = \log(\pi) \Rightarrow \pi = \exp(\theta) \]

\[ \phi = 1 \]

\[ a(\phi) = 1 \]

\[ c(y, \phi) = 0 \]

\[ b(\theta) = -\log(1 - \pi) = -\log \left(1 - \exp^{\theta}\right) \]

(b) (6 pts) State \( E\{Y\}, \ Var\{Y\} \) and the canonical link.

\textbf{Answer:}

\[ E\{Y\} = \frac{\partial b(\theta)}{\partial \theta} = \frac{e^{\theta}}{1 - e^{\theta}} = \frac{\pi}{1 - \pi} = \text{notation} = \mu \]

\[ Var\{Y\} = \frac{\partial^2 b(\theta)}{\partial \theta^2}a(\phi) = \frac{e^{\theta}}{(1 - e^{\theta})^2} = \frac{\pi}{(1 - \pi)^2} \]

Since

\[ \mu = \frac{e^{\theta}}{1 - e^{\theta}} \]

then the canonical link is

\[ \theta = \log \left(\frac{\mu}{1 + \mu}\right) \]
(c) Consider a two-arm randomized clinical trial where we study remission of a disease. The patients are randomly assigned to a treatment (arm 1) or to a control (arm 2). The protocol of the trial says that each arm stops whenever a remission is observed.

i. (6 pts) Specify a probability model that would be appropriate for this study. Use the canonical link function.

**Answer:**
The number of patients seen until the first remission follows a geometric distribution, with probability of an event different between the arms. Since the canonical link is

$$\log \left( \frac{\mu}{1 + \mu} \right) = \log \left( \frac{\pi / (1 - \pi)}{1 / (1 - \pi)} \right) = \log \pi^{\text{model}} = \beta_0 + \beta_1 \text{trt}$$

where trt takes value of 1 for treatment, and 0 otherwise.

ii. (6 pts) Suppose that the control arm stops at the fifth patient, and the treatment arm stops at the third patient. Write the log-likelihood function that will be maximized to obtain parameter estimates.

**Answer:**
The observed sequences of events are \(\bar{R}, \bar{R}, \bar{R}, \bar{R}, R\) for the control, and \(\bar{R}, \bar{R}, R\) for the treatment. Using the exponential family representation in 3(a), the log-likelihood is

$$l(\beta_0, \beta_1) = 4 \beta_0 + \log(1 - e^{\beta_0}) + \left[ 2 (\beta_0 + \beta_1) + \log(1 - e^{\beta_0 + \beta_1}) \right]$$
4. Researchers conducted a study to determine whether female breast cancer patients can be more accurately classified into subtypes using immunohistochemical (IH) examination. The study reported survival times (in months) of patients with negative and with positive outcomes of the exam. The data are shown below, where ’+‘ indicates a censored observation.

<table>
<thead>
<tr>
<th></th>
<th>Survival times $t_i$</th>
<th>n</th>
<th>$\sum_{i=1}^n t_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IH -</td>
<td>19 25 30 30+ 30+ 38 47 51 56 57 78 86 122+ 123+ 130+ 130+ 133+ 134+</td>
<td>18</td>
<td>1319</td>
</tr>
<tr>
<td>IH+</td>
<td>22 22+ 38 42 73 77 89 115 144+</td>
<td>9</td>
<td>622</td>
</tr>
</tbody>
</table>

(a) The researchers would like to estimate the probability of survival for each group, without making any parametric distributional assumptions.

i. (6 pts) State the assumptions of the Kaplan-Meier model for survival, and report $\hat{S}_{IH-}^{KM}(38)$.

**Answer:** The survival function is $S(t) = c_i$, $t_i \leq t < t_{i+1}$ for each observed event time $t_i$, where $1 \geq c_i \geq c_{i+1} \geq 0$.

Its Kaplan-Meier estimator is

$$\hat{S}_{IH-}^{KM}(38) = \prod_{i: t_i \leq t} \frac{n_i - d_i}{n_i} = (1 - 1/18) \cdot (1 - 1/17) \cdot (1 - 1/16) \cdot (1 - 1/13) = 0.7692$$

ii. (6 pts) Report the 95% confidence interval of $\hat{S}_{IH-}^{KM}(38)$.

**Answer:**

$$Var\{\log \hat{S}_{IH-}^{KM}(38)\} = \sum_{i: t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} = \frac{1}{18 \cdot 17} + \frac{1}{17 \cdot 16} + \frac{1}{16 \cdot 15} + \frac{1}{13 \cdot 12} = 0.0175$$

Then the 95% confidence interval for $\log \hat{S}_{IH-}^{KM}(38)$ is

$$\log 0.7692 \pm 1.96 \cdot \sqrt{0.0175} \rightarrow [L, R] = [-0.5218, -0.002]$$

and the approximate 95% confidence interval for $\hat{S}_{IH-}^{KM}(38)$ is

$$[e^L, e^R] = [0.5934, 0.9970]$$
(b) The researchers now assume that the survival function for each group follows an exponential distribution.

i. (6 pts) State the assumptions of the model, and report $S_{IH-}^{\text{exponential}}(38)$.

**Answer:**
The survival function is

$$S_{\text{placebo}}(t) = e^{-\lambda_{IH-} t}, \quad S_{\text{trt}}(t) = e^{-\lambda_{IH+} t},$$

where $\lambda$ is the reciprocal of the expected survival time. The MLE of $\lambda_{IH-}$ is

$$\hat{\lambda}_{IH-} = \frac{\sum_{i=1}^{18} \delta_i}{\sum_{i=1}^{18} t_i} = \frac{10}{1319} = 0.0075,$$

and

$$S_{IH-}^{\text{exponential}}(38) = e^{-0.0075 \cdot 38} = 0.7520,$$

comparable to the K-M estimate.

ii. (6 pts) Report the 95% confidence interval of $S_{IH-}^{\text{exponential}}(38)$. How does this confidence interval compare to the confidence interval in (a, ii)? Explain. (*Hint:* use the observed information to estimate $\text{Var}(S_{IH-}^{\text{exponential}}(38))$)

**Answer:**

$$SE\{\hat{\lambda}_{IH-}\} = \sqrt{\frac{\sum_{i=1}^{18} \delta_i}{\left(\sum_{i=1}^{18} t_i\right)^2}} = \sqrt{\frac{10}{1319^2}} = 0.0023$$

$$SE[\log S(t)] = \sqrt{\text{Var}(\log S(t))} = \sqrt{\text{Var}\{-\lambda t\}} = \sqrt{\left(t^2 \text{Var}(\lambda)\right)} = tSE(\lambda) = t \cdot 0.0023$$

Then the 95% confidence interval for $\log S_{IH-}^{\text{exponential}}(38)$ is

$$\log 0.7520 \pm 1.96 \cdot 0.0023 \cdot 38 \rightarrow [L, R] = [-0.4563, -0.1137]$$

and the approximate 95% confidence interval for $S_{IH-}^{\text{exponential}}(38)$ is

$$[e^L, e^R] = [0.6336, 0.8925]$$

The confidence interval is narrower due to the parametric nature of the estimation.
To determine the effectiveness of the immunohistochemical examination, the researchers fit the model given in the partial output below. In the output, the variable $i_h = 0$ for the IH negative and $i_h = 1$ for the IH positive group.

Call:
```r
survreg(formula = Surv(time, delta) ~ i_h, data = X, dist = "exponential")
```

Value Std. Error  z  p
(Intercept) 4.882 0.316 15.438 9.03e-54
i_h -0.395 0.493 -0.802 4.23e-01

Scale fixed at 1

Exponential distribution
Loglik(model)= -97.2  Loglik(intercept only)= -97.5
Chisq= 0.62 on 1 degrees of freedom, p= 0.43

i. (6 pts) State the assumptions of the model, and interpret the parameters.

Answer:

This is an accelerated failure time model, which assumes that the covariates have the effect of adjusting time to event

$$ S(t) = S_0(t \cdot e^{\beta_0 + \beta_1 \cdot i_h}) $$

where

$S_0(t)$ is the survival function of the standard exponential random variable.

ii. (6 pts) Report the estimate of $\hat{S}_{IH-}(38)$ according to this model. Does the estimate agree with the result in (b)? Explain.

Answer:

$$ \hat{S}_{AF}^{IH-}(38) = \exp(-38 \cdot e^{-4.882}) = 0.7496 $$

Up to numeric roundings, $\hat{S}_{IH-}^{AF}(38) = \hat{S}_{IH-}^{exponential}(38)$. The two models specify the same likelihood functions for each of the groups, and therefore yield the same parameter estimates.
5. Consider a proportional odds model for an ordered multinomial response $Y$, as function of the predictor variable $X$. For each question below, circle TRUE or FALSE, and provide the rationale.

(a) (6 pts) Suppose that the cumulative odds ratio of a particular category of $Y$ for two levels of $X$ is 3. If we invert the order of the response categories, then the odds ratio will equal to $1/3$.

**TRUE**      **FALSE**

Answer:
True. The model is defined as

$$\log\left( \frac{P\{Y \leq j\}}{1 - P\{Y \leq j\}} \right) = \alpha_j + \beta X$$

The model in the new order of the response categories can be expressed in terms of the old order of the response categories as

$$-\log\left( \frac{P\{Y \leq j\}}{1 - P\{Y \leq j\}} \right) = \alpha_j + \beta X, \text{ i.e.}$$

$$\log\left( \frac{P\{Y \leq j\}}{1 - P\{Y \leq j\}} \right) = -\alpha_j - \beta X$$

and the odds ratio

$$OR = \frac{\text{odds } P\{Y \leq j \mid X + 1\}}{\text{odds } P\{Y \leq j \mid X\}} = e^{-\beta} = \frac{1}{\beta}$$

(b) (6 pts) A positive value of the parameter $\beta$ associated with the predictor $X$ can be interpreted as the positive association between the predictor and the latent continuous variable underlying the response.

**TRUE**      **FALSE**

Answer:
False. A positive parameter indicates a negative association. The model expresses the fact that an increase in $X$ is associated with a higher odds of lower categories of $Y$. 