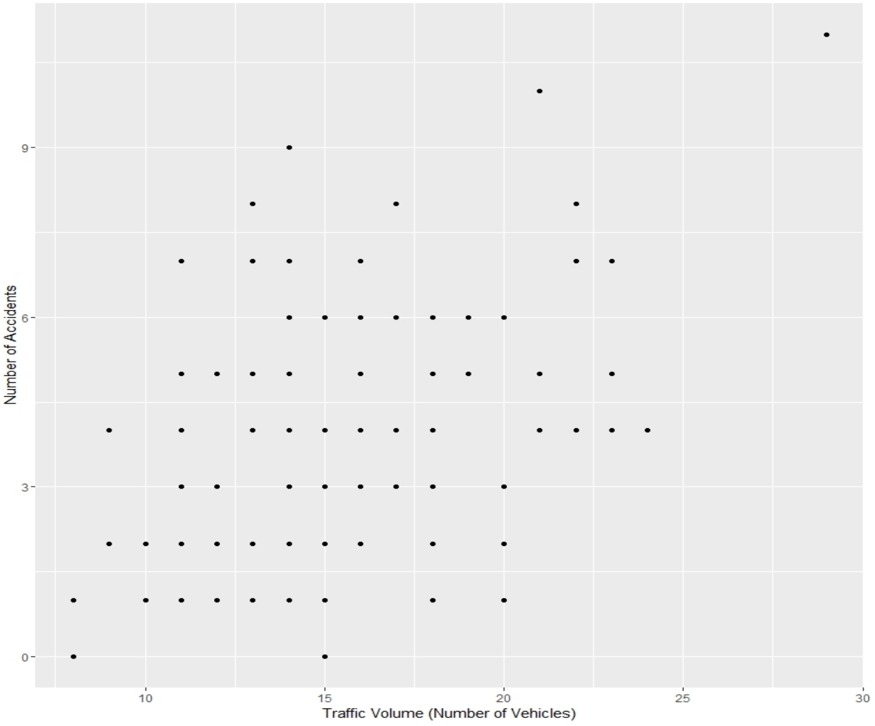
Name:

This midterm has one short answer section and one multiple choice section. Each section contains multiple sub-questions. If a multiple choice question has more than one correct answer, circle all correct answers. The questions are on pages 2-14 of this packet. Please let me know if you have any questions.

Short Answer

Your textbook considers a problem in which the objective is to model the association between the average number of accidents at intersections and the traffic volume. This question will revisit that problem. The figure below shows a scatterplot of the number of accidents plotted against the traffic volume for 100 randomly selected intersections. In this exercise, we will consider a generalized linear model for these data with a Poisson random component. We will define the systematic component by $\mu_i = x_i\beta$, where x_i is the traffic volume at intersection i, μ_i is the mean number of accidents at intersection i, and β is an unknown parameter.



1. Define appropriate random variables for this problem. State the support of the random variables that you define.

   Let Y_i be a random variable associated with the number of accidents at intersection i, where $i = 1, \ldots, n$. The support of Y_i is the set of positive integers, $\{0, 1, 2, \ldots\}$.

2. Express the probability mass function for one of the random variables that you defined in exponential dispersion form (or natural exponential form, if appropriate) as a function of x_i and β.

$$\begin{aligned} P(Y_i = y_i \mid \mu_i) &= \frac{\mu_i^{y_i} \exp(-\mu_i)}{y_i!}, \quad y_i = 0, 1, 2, \ldots, \\ &= \exp\left[y_i \log(\mu_i) - \mu_i - \log(y_i!)\right] \\ &= \exp\left[y_i \log((x_i\beta)) - x_i\beta - \log(y_i!)\right] \end{aligned}$$

3. Comment briefly on the mechanism of interest, and state which parameter (or parameters) captures this mechanism.

   In this model, the mean number of traffic accidents increases in proportion to the traffic volume, and the parameter β is the proportionality constant.

4. The output below contains estimates of the parameters of the generalized linear model with the Poisson random component and $\mu_i = x_i\beta$. Use the output below to answer (a), (b), and (c) on the next page.

(a) Construct a 95% Wald interval for β.

$$0.2581 \pm (1.96)0.0129 = [0.233, 0.283]$$

(b) Provide an estimate and standard error for $V\{Y_i \mid x_i = 2\}$, where Y_i is the random variable that you defined in question 1 above. (The notation $V\{Y_i \mid x_i = 2\}$ means "the variance of Y_i when x_i is equal to 2.")

$$\begin{aligned}
2\beta &= V\{Y_i \mid x_i = 2\} \\
2\hat{\beta} &= 0.5162 \\
SE\{2\hat{\beta}\} &= 2SE\{\hat{\beta}\} = 2(0.0129) = 0.0258
\end{aligned}$$

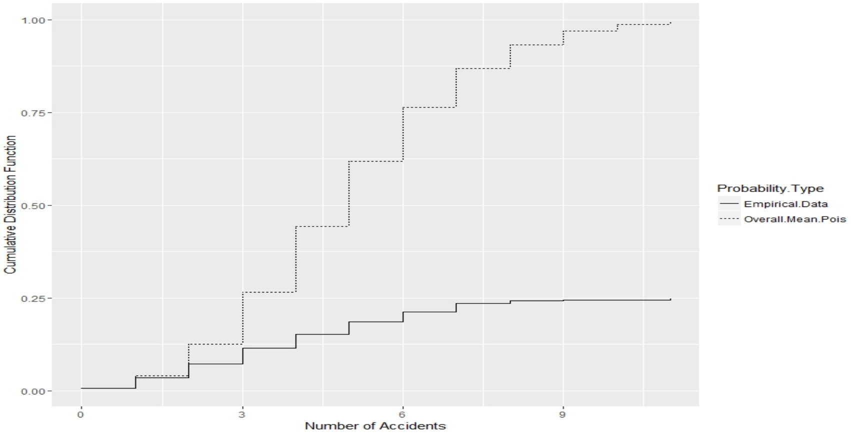
(c) Provide an estimate and standard error for the probability of at least one traffic accident if $x_i = 2$.

$$\begin{aligned}
\hat{P}(Y_i \geq 1 \mid x_i = 2) &= 1 - \hat{P}(Y_i = 0 \mid x_i = 1) \\
&= 1 - \frac{(2\hat{\beta})^0 \exp(-2\hat{\beta})}{0!} \\
&= 1 - \exp(-2\hat{\beta}) = 1 - \exp(-2(0.2581)) = 0.4032.
\end{aligned}$$

By the delta method,

$$\begin{aligned}
SE\{1 - \exp(-2\hat{\beta})\} &= 2\exp(-2\hat{\beta})SE\{\hat{\beta}\} \\
&= 2\exp(-2(0.2581))(0.0129) = 1.19(0.0129) = 0.015.
\end{aligned}$$

5. The graph below compares two cumulative distribution functions. The solid line shows the empirical cumulative distribution function based on the data. The height of the solid line gives the proportion of intersections in the data set where the number of accidents is less than or equal to the value on the horizontal axis. The dashed line is the cumulative distribution function of a single Poisson distribution with a mean of 4. Note that 4 is the average number of accidents in the data set. Observe that the two cumulative distribution functions are obviously completely different. Does this mean that the Poisson distribution is an inappropriate random component for these data? Explain why or why not.



No – The empirical CDF in the plot gives a picture of the marginal distribution of the data. The random component of the generalized linear model specifies a conditional distribution at each level of the traffic volume. The Poisson distribution with a constant mean equal to the average number of traffic accidents in the data set is not the random component that we specify in our model.

6. For the particular generalized linear model considered in this exercise, the maximum likelihood estimator of β has a closed form expression. Give a closed form mathematical expression for the maximum likelihood estimator of β.

   The log likelihood is

   $$\ell(\beta) = \left\{ \sum_{i=1}^{n} [y_i \log(x_i\beta) - x_i\beta - \log(y_i!)] \right\}$$

   $$\propto n\bar{y}_n \log(\beta) - n\bar{x}_n\beta.$$

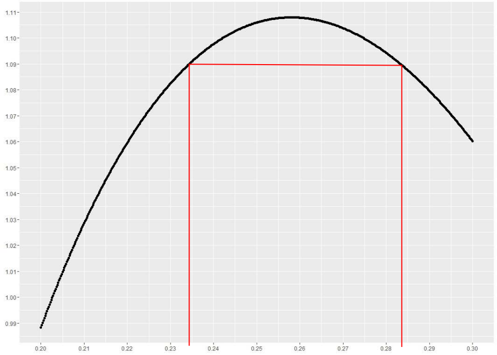
   The score equation is

   $$S(\beta) = n\bar{y}_n/\beta - n\bar{x}_n.$$

   Setting the score equation equal to zero and solving for β gives $\hat{\beta} = \bar{x}_n^{-1}\bar{y}_n$.

   To confirm that $\hat{\beta}$ is a maximum, note that the derivative of $S(\beta)$ is $S'(\beta) = -n\bar{y}_n/\beta^2$, which is strictly negative.

7. **Extra Credit**: The figure below is a plot of the log density function (natural logarithm of density function) of a gamma distribution with a shape parameter of 5 and a rate parameter of 15.5. Sketch a 95% likelihood ratio interval for β on the graph below. (Note that 15.5 is the average traffic volume in the data set.)



By definition, the likelihood ratio interval is

$$C = \{\beta_0 : -2[\ell(\beta_0) - \ell(\hat{\beta})] < \chi_1^2(0.95)\}.$$

By exercise 6,

$$-2[\ell(\beta_0) - \ell(\hat{\beta})] = -2[n\bar{y}_n\log(\beta) - n\bar{x}_n\beta - n\bar{y}_n\log(\hat{\beta}) + n\bar{x}_n\hat{\beta}].$$

Therefore, $\beta \in C \leftrightarrow$

$$\begin{aligned} \bar{y}_n\log(\beta) - \bar{x}_n\beta - \bar{y}_n\log(\hat{\beta}) + \bar{x}_n\hat{\beta} > -\frac{\chi_1^2(0.95)}{2n} \leftrightarrow \\ \beta^{\bar{y}_n}\exp(-\bar{x}_n\beta) > \exp(-\frac{\chi_1^2(0.95)}{2n} + \bar{y}_n\log(\hat{\beta}) - \bar{x}_n\hat{\beta}) \leftrightarrow \\ \frac{\bar{x}_n^{\bar{y}_n+1}\beta^{\bar{y}_n}\exp(-\bar{x}_n\beta)}{\Gamma(\bar{y}_n+1)} > \exp(-\frac{\chi_1^2(0.95)}{2n} - \bar{x}_n\hat{\beta})\frac{\bar{x}_n^{\bar{y}_n+1}\hat{\beta}^{\bar{y}_n}}{\Gamma(\bar{y}_n+1)}. \end{aligned}$$

The left side of the final expression above is a pdf of a gamma distribution with shape $\bar{y}_n + 1$ and rate $\bar{x}_n$. Denote the pdf of a gamma distribution with shape $\bar{y}_n + 1 =$ and rate $\bar{x}_n = \bar{y}/\hat{\beta} = 4/0.2581 = 15.5$

by $f_\beta(\beta \mid 5, 15.5)$. Then, $\beta \in C \leftrightarrow$

$$\begin{aligned}
\log(f_\beta(\beta \mid 5, 15.5)) &\geq -\frac{1.96^2}{200} + 5\log(15.5) - \log(\Gamma(5)) - 0.2581(15.5) + (4)\log(0.2581) \\
&\geq -\frac{1.96^2}{200} + 5\log(15.5) - \log(\Gamma(5)) - 0.2581(15.5) + (4)\log(0.2581) \\
&= -0.019208 + 13.7042 - \log(4!) - 0.2581(15.5) + (4)\log(0.2581) = 1.09
\end{aligned}$$

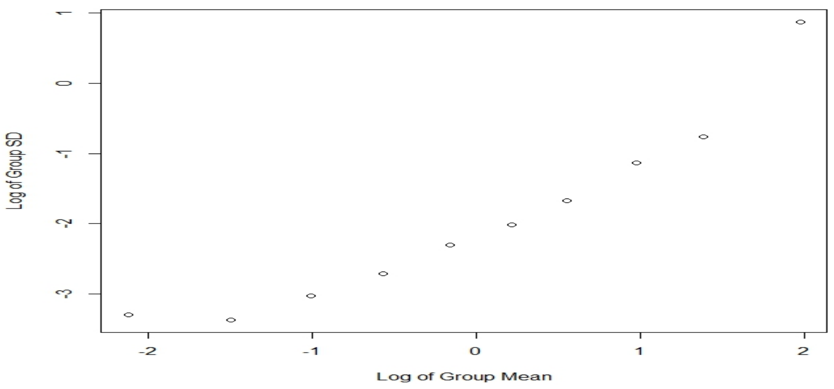
The graph above shows a plot of $\log(f_\beta(\beta \mid 5, 15.5))$ on the vertical axis against β on the horizontal axis. A likelihood ratio confidence interval for β is the set of β such that curve showing $\log(f_\beta(\beta \mid 5, 15.5))$ exceeds 1.09. As the red lines in the graph above indicate, this corresponds approximately to an interval of $[0.235, 0.284]$. The likelihood ratio interval is quite similar to the Wald interval calculated in part 4a above.

Multiple Choice

For each of the following questions, circle the correct answer or answers. Each question has at least one correct answer. Circle all answers that are correct.

1. A researcher would like to understand the association between income and several socioeconomic variables. A random sample of $n = 1000$ adults is selected. Associate random variables $Y_1, \ldots, Y_n$ with the incomes of the n adults. Let $E[Y_i \mid \boldsymbol{x}_i] = \mu_i$, where $\boldsymbol{x}_i$ is the vector of socioeconomic variables in the population. In preparation for selecting an appropriate random component, the researcher constructs the Box-Cox plot shown in the figure below, where logarithms of group standard

deviations are plotted on the vertical axis with the logarithms of group means on the horizontal axis. Note that the ordinary least squares slope of the regression of the log group standard deviations on the log group means is 0.97. Which of the following are definitely true?



(a.) A gamma random component seems reasonable because for a gamma random component, $V\{Y_i \mid \boldsymbol{x}_i\} \propto \mu_i^\theta$ with $\theta = 1$, and the ordinary least squares slope suggests a value of $\theta = 1$.

(b.) An inverse-Gaussian random component seems reasonable because for an inverse-Gaussian random component, $V\{Y_i \mid \boldsymbol{x}_i\} \propto \mu_i^\theta$ with $\theta = 2$, and the ordinary least squares slope suggests a value of $\theta = 2$.

(c.) A gamma random component seems reasonable because the canonical link for a gamma distribution is is $g(\mu) \propto \mu^{-1}$, and the ordinary least squares slope suggests this canonical link function.

(d.) A gamma random component seems reasonable because for a gamma distribution, $V\{Y_i \mid \boldsymbol{x}_i\} \propto \mu_i^\theta$ with $\theta = 2$, and the ordinary least squares slope suggests a value of $\theta = 2$.

(e.) An inverse-Gaussian random component seems reasonable because the canonical link for an inverse-Gaussian distribution is $g(\mu) \propto \mu^{-2}$, and the ordinary least squares slope suggests this canonical link function.

2. The following are five statements about the normal distributions. Which are definitely true?

   (a.) We should never use the normal distribution if our data are discrete.

   (b.) Limitations of the normal distribution for modeling include the unrestricted sample space and fixed shape.

   (c.) One of the reasons that the normal distribution is widely used for inference is that estimators often have asymptotically normal distributions due to the Central Limit Theorem.

   (d.) The normal distribution is **NOT** a member of the exponential dispersion family.

   (e.) The normal distribution is **NOT** a location/scale family.

3. Let Y have pdf/pmf given by

$$f_Y(y \mid \theta_1, \theta_2) = \exp\left[\theta_1 T_1(y) + \theta_2 T_2(y) - B(\theta_1, \theta_2) + c(y)\right].$$

Which of the following are definitely true?

(a.) $T_1(Y)$ is a sufficient statistic for θ_1.

(b.) Y has a distribution in the natural exponential family.

(c.) Y is a sufficient statistic for θ_2.

(d.) $E[T_2(Y)] = \frac{\partial B(\theta_1,\theta_2)}{\partial \theta_2}$

(e.) $f_Y(y \mid \theta_1, \theta_2)$ is NOT written in a canonical form.

(f.) $E[Y] = \frac{\partial B(\theta_1,\theta_2)}{\partial \theta_2}$.

4. Let Y have pdf/pmf given by

$$f_Y(y \mid \theta_1, \theta_2) = \exp\left[\phi\{y\theta - b(\theta)\} + c(y, \phi)\right].$$

Let $E[Y] = \mu$. Which of the following are definitely true?

(a.) Y is a member of the natural exponential family.

(b.) $V\{Y\} = \phi^{-1} b''(\theta)$

(c.) $V\{Y\} = \phi b'(\theta)$

(d.) $\mu = b''(\theta)$

(e.) $\mu = b'(\theta)$

(f.) $\mu = \phi b''(\theta)$

5. Let $Y_1, \ldots, Y_n$ be a sample from a generalized linear model such that the pdf/pmf of Y_i is

$$f(y_i \mid x_i) = \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right],$$

and $g(E[Y_i \mid x_i]) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $g(\cdot)$ is the canonical link function. Which is the following are definitely true?

(a.) Computing the maximum likelihood estimator of $\boldsymbol{\beta}$ requires an estimator of ϕ.

(b.) We can obtain the maximum likelihood estimator of $\boldsymbol{\beta}$ without knowledge of ϕ.

(c.) The score function for estimating $\boldsymbol{\beta}$ is of the form

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}))\boldsymbol{x}_i.$$

(d.) The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\boldsymbol{I}^{0.5}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{I}),$$

where

$$\boldsymbol{I} = \phi \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' V(\mu_i),$$

and $V(\mu_i) = b''(b'^{-1}(\mu_i))$.

(e.) The canonical link function for this model satisfies $g(b'(\theta_i)) = \theta_i$

(f.) The canonical link function for this model satisfies $g(\mu_i) = V^{-1}(\mu_i)$.