1. (a) For $y > 0$, the density of exponential distribution with mean μ can be written as

$$
\begin{aligned}
f(y) = \frac{1}{\mu} e^{-\frac{y}{\mu}} &= \exp\left\{ -\frac{y}{\mu} + \log\left(\frac{1}{\mu}\right) \right\} \\
&= \exp\left\{ y\theta + \log(-\theta) \right\} && (\text{where } \theta = -\mu^{-1}) \\
&= \exp\left\{ \frac{y\theta - A(\theta)}{\phi} + c(\theta, \phi) \right\},
\end{aligned}
$$

where $A(\theta) = -\log(-\theta)$, $\phi = 1$ and $c(\theta, \phi) = 0$. Then $\mu = A'(\theta) = -\frac{1}{\theta}$ and hence the canonical link function is $\theta = \psi(\mu) = -\frac{1}{\mu}$.

(b) Let n denote the number of rows in the design matrix $\mathbf{X}$. The log-likelihood and its first partial derivative w.r.t. the linear predictors $\eta_i := \mathbf{x}_i^\top \boldsymbol{\beta}$ is

$$
l(\boldsymbol{\eta}|\mathbf{y}) = \sum_{i=1}^{n} \left(-\eta_i - y_i e^{-\eta_i}\right), \quad \frac{\partial l(\boldsymbol{\eta})}{\partial \eta_i} = y_i e^{-\eta_i} - 1.
$$

Hence, by chain rule, we have

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} (y_i e^{-\eta_i} - 1) x_{ij}, \\
\frac{\partial}{\partial \beta_k} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \left(\frac{\partial}{\partial \eta_i} \frac{\partial l}{\partial \beta_j} \right) \frac{\partial \eta_i}{\partial \beta_k} = -\sum_{i=1}^{n} y_i e^{-\eta_i} x_{ij} x_{ik}
\end{aligned}
$$

for each j, k. Summarizing all the results gives us

$$
\nabla l = \sum_{i=1}^{n} (y_i e^{-\mathbf{x}_i^\top \boldsymbol{\beta}} - 1) \mathbf{x}_i, \quad \mathbf{H} = -\sum_{i=1}^{n} y_i e^{-\mathbf{x}_i^\top \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}_i^\top.
$$

2. Let n and p denote, respectively, the number of rows and columns in the design matrix $\mathbf{X}$.

(a) Since $Y_i \sim \text{Degenerate}(0)$ w.p. p_i and $Y_i \sim \text{Po}(\lambda_i)$ w.p. $1 - p_i$, by law of total probability, we have

$$
\begin{aligned}
\mathbb{P}(Y_i = 0|\mathbf{X}_i) &= \mathbb{P}(Y_i \sim 0)\mathbb{P}(Y_i = 0|Y_i \sim 0, \mathbf{X}_i) + \mathbb{P}(Y_i \sim \text{Po}(\lambda_i))\mathbb{P}(Y_i = 0|Y_i \sim \text{Po}(\lambda_i), \mathbf{X}_i) \\
&= p_i \cdot 1 + (1 - p_i) e^{-\lambda_i}, \\
\mathbb{P}(Y_i = k|\mathbf{X}_i) &= p_i \cdot 0 + (1 - p_i) \frac{e^{-\lambda_i} \lambda_i^k}{k!}
\end{aligned}
$$

for each i, where $k \in \{1, 2, \cdots\}$. To be more concise, we may write

$$
f_i(k) := \mathbb{P}(Y_i = k|\mathbf{X}_i) = (1 - p_i) \frac{e^{-\lambda_i} \lambda_i^k}{k!} \left(1 + \mathbb{1}_{\{0\}}(k) \frac{p_i e^{\lambda_i}}{1 - p_i}\right)
$$

for each $i \in \{1, \cdots, n\}$, where $k \in \{0, 1, 2, \cdots\}$.

(b) The likelihood function w.r.t. the parameters p_i and λ_i is

$$
L(\mathbf{p}, \boldsymbol{\lambda}|\mathbf{y}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} (1 - p_i) \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \left(1 + \mathbb{1}_{\{0\}}(y_i) \frac{p_i e^{\lambda_i}}{1 - p_i}\right),
$$

which can also be re-parameterized in terms of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ after straightforward substitutions, but here we keep the likelihood in terms of p_i and λ_i to make our later analysis more convenient. For the same reason, we will facilitate the analysis base on the log-likelihood

$$l(\mathbf{p}, \boldsymbol{\lambda}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ \log(1-p_i) - \lambda_i + y_i \log(\lambda_i) - \log(y_i!) + \mathbb{1}_{\{0\}}(y_i) \log\left(1 + \frac{p_i e^{\lambda_i}}{1-p_i}\right) \right\}.$$

(c) For each i, denote $\eta_i := \mathbf{x}_i^\top \boldsymbol{\beta} = \log\left(\frac{p_i}{1-p_i}\right)$ and $\zeta_i := \mathbf{x}_i^\top \boldsymbol{\gamma} = \log(\lambda_i)$. Then it follows that

$$\begin{aligned}
\frac{\partial l}{\partial p_i} &= \frac{-1}{1-p_i} + \mathbb{1}_{\{0\}}(y_i) \left(\frac{e^{\lambda_i}(1-p_i)^{-2}}{1+e^{\lambda_i+\eta_i}} \right), & \frac{\mathrm{d}p_i}{\mathrm{d}\eta_i} &= p_i(1-p_i), \\
\frac{\partial l}{\partial \lambda_i} &= -1 + \frac{y_i}{\lambda_i} + \mathbb{1}_{\{0\}}(y_i) \left(\frac{e^{\lambda_i+\eta_i}}{1+e^{\lambda_i+\eta_i}} \right), & \frac{\mathrm{d}\lambda_i}{\mathrm{d}\zeta_i} &= \lambda_i, \\
\frac{\partial \eta_i}{\partial \beta_j} &= \frac{\partial \zeta_i}{\partial \gamma_j} = x_{ij} \quad \text{for each } j \in \{1, \cdots, p\}.
\end{aligned}$$

By chain rule, we have

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l}{\partial p_i} \frac{\mathrm{d}p_i}{\mathrm{d}\eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \frac{-1}{1+e^{-\eta_i}} + \frac{\mathbb{1}_{\{0\}}(y_i)}{1+e^{-(e^{\zeta_i}+\eta_i)}} \right\} x_{ij}, \\
\frac{\partial l}{\partial \gamma_j} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \lambda_i} \frac{\mathrm{d}\lambda_i}{\mathrm{d}\zeta_i} \frac{\partial \zeta_i}{\partial \gamma_j} = \sum_{i=1}^{n} \left\{ -e^{\zeta_i} + y_i + \frac{\mathbb{1}_{\{0\}}(y_i) e^{\zeta_i}}{1+e^{-(e^{\zeta_i}+\eta_i)}} \right\} x_{ij}
\end{aligned}$$

for each $j \in \{1, \cdots, p\}$, so the gradient w.r.t. $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ is

$$\nabla l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \left(\frac{\partial l}{\partial \beta_1}, \cdots, \frac{\partial l}{\partial \beta_p}, \frac{\partial l}{\partial \gamma_1}, \cdots, \frac{\partial l}{\partial \gamma_p} \right)^\top.$$

Now, apply chain rule again on each of the partial derivatives. For $j, k \in \{1, \cdots, p\}$,

$$\begin{aligned}
\frac{\partial}{\partial \beta_k} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \left(\frac{\partial}{\partial \eta_i} \frac{\partial l}{\partial \beta_j} \right) \frac{\partial \eta_i}{\partial \beta_k} \\
&= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \eta_i} \sum_{q=1}^{n} \left(\frac{-1}{1+e^{-\eta_q}} + \frac{\mathbb{1}_{\{0\}}(y_q)}{1+e^{-(e^{\zeta_q}+\eta_q)}} \right) x_{qj} \right\} \frac{\partial \eta_i}{\partial \beta_k} \\
&= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \eta_i} \left(\frac{-1}{1+e^{-\eta_i}} + \frac{\mathbb{1}_{\{0\}}(y_i)}{1+e^{-(e^{\zeta_i}+\eta_i)}} \right) x_{ij} \right\} x_{ik} \\
&= \sum_{i=1}^{n} \left\{ \frac{-e^{-\eta_i}}{(1+e^{-\eta_i})^2} + \frac{\mathbb{1}_{\{0\}}(y_i) e^{-(e^{\zeta_i}+\eta_i)}}{(1+e^{-(e^{\zeta_i}+\eta_i)})^2} \right\} x_{ij} x_{ik}, \\
\frac{\partial}{\partial \gamma_k} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \zeta_i} \left(\frac{-1}{1+e^{-\eta_i}} + \frac{\mathbb{1}_{\{0\}}(y_i)}{1+e^{-(e^{\zeta_i}+\eta_i)}} \right) x_{ij} \right\} x_{ik} \\
&= \sum_{i=1}^{n} \frac{\mathbb{1}_{\{0\}}(y_i) e^{\zeta_i - (e^{\zeta_i}+\eta_i)}}{(1+e^{-(e^{\zeta_i}+\eta_i)})^2} x_{ij} x_{ik}, \\
\frac{\partial}{\partial \gamma_k} \frac{\partial l}{\partial \gamma_j} &= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \zeta_i} \left(-e^{\zeta_i} + y_i + \frac{\mathbb{1}_{\{0\}}(y_i) e^{\zeta_i}}{1+e^{-(e^{\zeta_i}+\eta_i)}} \right) x_{ij} \right\} x_{ik} \\
&= \sum_{i=1}^{n} \left\{ -e^{\zeta_i} + \mathbb{1}_{\{0\}}(y_i) \frac{e^{\zeta_i}(1+e^{-(e^{\zeta_i}+\eta_i)} + e^{\zeta_i - (e^{\zeta_i}+\eta_i)})}{(1+e^{-(e^{\zeta_i}+\eta_i)})^2} \right\} x_{ij} x_{ik}.
\end{aligned}$$

Hence, finally, the hessian matrix can obtained as

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top} & \dfrac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\gamma}^\top} \\ \dfrac{\partial^2 l}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\beta}^\top} & \dfrac{\partial^2 l}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\gamma}^\top} \end{bmatrix},$$

where the block matrices are defined by

$$\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top}\right]_{jk} = \frac{\partial^2 l}{\partial \beta_j \partial \beta_k}, \quad \left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\gamma}^\top}\right]_{jk} = \left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\beta}^\top}\right]_{kj} = \frac{\partial^2 l}{\partial \beta_j \partial \gamma_k}, \quad \left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\gamma}^\top}\right]_{jk} = \frac{\partial^2 l}{\partial \gamma_j \partial \gamma_k}$$

for $j, k \in \{1, \cdots, p\}$.

(d) Notice that the naive poisson regression is a special case of the zero-inflated one, where the former one restricts $p_i = 0$ for each i, so the reduced model is only fitted with $\boldsymbol{\gamma}$. Hence, we may use the nested model test H_0 : naive poisson model against H_1 : otherwise. Find the MLE of both models and compare the observed ΔD with the χ^2-quantile, with the degree-of-freedom being $2p - p = p$. Rejection of H_0 indicates the under-fit of the naive Poisson regression.

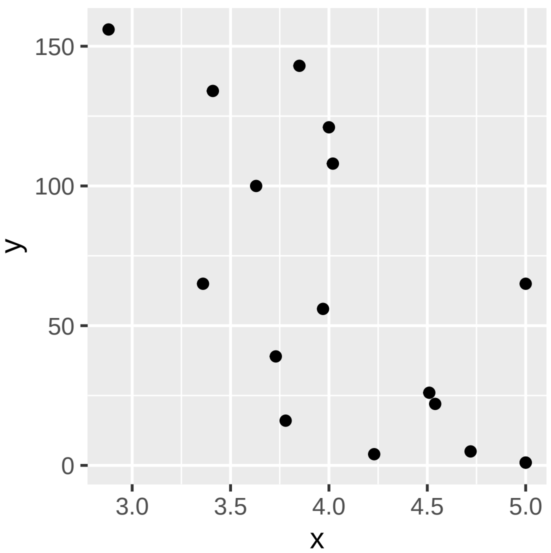
Remark. Open-ended.

3. In the given table, the first row should be y_i and the second row should be x_i.

```r
library(stats4)
library(ggplot2)

y <- c(65,156,100,134,16,108,121,
       4,39,143,56,26,22,1,1,5,65)
x <- c(3.36,2.88,3.63,3.41,3.78,4.02,4,4.23,
       3.73,3.85,3.97,4.51,4.54,5,5,4.72,5)
n <- length(x)
```

(a)
```r
df <- data.frame(x,y)
ggplot(df, aes(x=x,y=y)) + geom_point()
```



As observed, the y_i decreases in general when x_i increases.

(b) The model can be fitted with the MLE of (β_1, β_2), which can be obtained as follows:

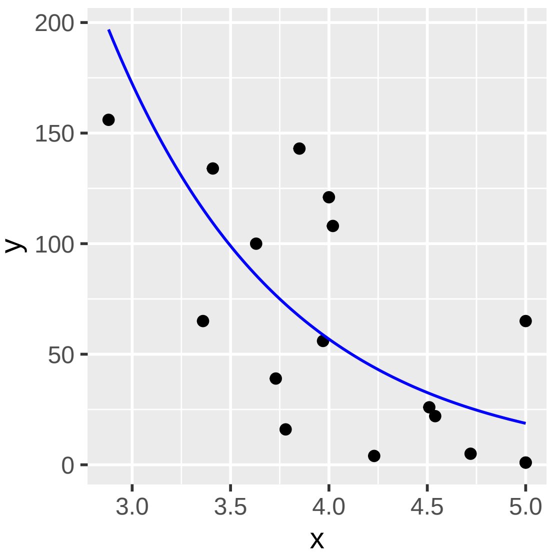
```
minusLogLik <- function(b1 = 0, b2 = 0)
    sum(y*exp(-b1-b2*x)+(b1+b2*x))

fit <- mle(minusLogLik)
b_hat <- coef(fit)
fitted_line <- function(x)
    exp(b_hat[1]+x*b_hat[2])

b_hat # MLE of (beta1, beta2)
```

```
##        b1        b2
##  8.477006 -1.109194
```

```
ggplot(df, aes(x=x,y=y)) +
            geom_point() +
            geom_function(fun = fitted_line, color="blue")
```



(c)
```
y_hat <- fitted_line(x)
r <- (y-y_hat)/y_hat
r
```

```
##  [1] -0.43770454 -0.20758666  0.16711794  0.22529837 -0.77945753  0.94271027
##  [7]  1.12880223 -0.90917513 -0.49142966  1.13023968 -0.04701411 -0.19462371
## [13] -0.29546970 -0.94665832 -0.94665832 -0.80449583  2.46720931
```

```
sum(r^2)/(n-2) # Compare with \phi (=1 from Q1)
```

```
## [1] 0.9388941
```

Since most of the r_i are not far from 0 and $\mathcal{X}^2/(n-p)$ is close to $\phi = 1$, we consider the model above as a good fit. (Concretely, we can check through the Pearson's GoF test, in which we cannot reject the null hypothesis base on the observed $\mathcal{X}^2$.)

Remark. Open-ended.

4. The log-likelihoods for the ungrouped and grouped model w.r.t. $\boldsymbol{\pi}$ are, respectively,

$$\begin{aligned} l(\boldsymbol{\pi}|z_{ij}) &= \log \prod_{i\in I}\prod_{j=1}^{m_i} \pi_i^{z_{ij}}(1-\pi_i)^{1-z_{ij}} \\ &= \sum_{i\in I}\sum_{j=1}^{m_i} \left(z_{ij}\log(\pi_i) + (1-z_{ij})\log(1-\pi_i)\right), \\ l(\boldsymbol{\pi}|y_i) &= \log \prod_{i\in I} \binom{m_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{m_i-y_i} \\ &= \sum_{i\in I}\left(\log\binom{m_i}{y_i} + y_i\log(\pi_i) + (m_i-y_i)\log(1-\pi_i)\right). \end{aligned}$$

Now, notice that for all $k \in I$,

$$\frac{\partial l(\boldsymbol{\pi}|z_{ij})}{\partial \pi_k} = \frac{\sum_{j=1}^{m_k} z_{kj}}{\pi_k} - \frac{\sum_{j=1}^{m_k}(1-z_{kj})}{1-\pi_k} = \frac{y_k}{\pi_k} - \frac{m_k - y_k}{1-\pi_k} = \frac{\partial l(\boldsymbol{\pi}|y_i)}{\partial \pi_k}$$

and the canonical link function for π_i in both models is the logit function, so their partial derivatives w.r.t. the GLM parameters $\boldsymbol{\beta}$ will match as well. Therefore, by chain rule, we can use the above partial derivatives to obtain $\nabla l(\boldsymbol{\beta}|z_{ij})$ and $\nabla l(\boldsymbol{\beta}|y_i)$, and the above arguments showed that they are equal.

5. The additional regularization term should be $-\boldsymbol{\beta}^\top\boldsymbol{\beta}$ if the objective is to maximize.

   By linearity of differentiation, we only need to further focus on the derivatives of the additional term $-\boldsymbol{\beta}^\top\boldsymbol{\beta}$ in the log-likelihood, and the rest are the same as what has been derived in class (page 53-54 of Lecture Note 9).

   The first partial derivative of the log-likelihood w.r.t. β_i involves an additional $-2\beta_i$, and the second partial derivative w.r.t. β_i, β_j involves an additional -2 if $i = j$ and involves nothing else if $i \neq j$. Therefore, we have

$$\nabla l(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} \frac{y_i - \mu_i}{g'(\mu_i)V(\mu_i)}\mathbf{x}_i - 2\boldsymbol{\beta}, \quad \mathbf{H}_{jk} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{g'(\mu_i)}\frac{\mathrm{d}}{\mathrm{d}\mu_i}\left(\frac{y_i-\mu_i}{g'(\mu_i)V(\mu_i)}\right) - 2\mathbb{1}\{j=k\}.$$

   The non-linear equations for solving $\boldsymbol{\beta}$ is the first-order condition $\nabla l(\boldsymbol{\beta}) = \mathbf{0}$.

6. Recall that the deviance is given by $D = 2[l(\mathbf{y}|\mathbf{X},\mathbf{y}) - l(\hat{\boldsymbol{\mu}}|\mathbf{X},\mathbf{y})]$, where $\hat{\boldsymbol{\mu}}$ is the MLE of $\boldsymbol{\mu}$ under the GLM setting (i.e. μ_i is related to the linear combination of independent variables through a link function, so its MLE is obtained from the MLE of those coefficients and the invariance property).

   (a) The model assumption is $Y_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(\mu_i, \sigma^2)$, where $\mu_i = \mathbf{x}_i^\top\boldsymbol{\beta}$. Then, the log-likelihood is

$$l(\boldsymbol{\mu}|\mathbf{X},\mathbf{y}) = \sum_{i=1}^{n}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i-\mu_i)^2}{2\sigma^2}\right],$$

   indicating that

$$D = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})^2 = \frac{1}{\sigma^2}\mathbf{y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y}.$$

   The last equality used the fact that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

(b) We have $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Bernoulli}(\mu_i)$, where $\mu_i = \frac{1}{1+\exp\{-\mathbf{x}_i^\top \boldsymbol{\beta}\}}$. Then, the log-likelihood is

$$l(\boldsymbol{\mu}|\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \log(1-\mu_i) \right],$$

indicating that

$$\begin{aligned} D &= 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{1-y_i}\right) + \log(1-y_i) - y_i \log\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) - \log(1-\hat{\mu}_i) \right] \\ &= 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{1-y_i}\right) + \log(1-y_i) - y_i \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \log(1+e^{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}) \right]. \end{aligned}$$

Here, the MLE $\hat{\boldsymbol{\beta}}$ does not have a close-form in general.

(c) We have $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Po}(\mu_i)$, $\mu_i = \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}$. Thus

$$l(\boldsymbol{\mu}|\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \left[-\mu_i + y_i \log(\mu_i) - \log(y_i!) \right],$$

indicating that

$$D = 2\sum_{i=1}^{n} \left[e^{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}} - y_i \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - y_i + y_i \log(y_i) \right].$$

Here, the MLE $\hat{\boldsymbol{\beta}}$ does not have a close-form in general.