1. Let $y_i$ be realizations of independent random variables $Y_i$ with Poisson $(\mu_i)$ distributions, where $E(Y_i) = \mu_i$, for $i = 1, \ldots, n$.

(a) Obtain the expression for the deviance for comparison of the full model, which assumes a different $\mu_i$ for each $y_i$, with a reduced model defined by a Poisson GLM with link function $g(\cdot)$. That is, under the reduced model, $g(\mu_i) = \eta_i = x_i^T \beta$, where $\beta = (\beta_1, \ldots, \beta_p^T)$ (with $p < n$ is the vector of regression coefficients corresponding to covariates $x_i = (x_{i1}, \ldots, x_{ip})^T$.

Solution:

By definition, the deviance is

$$D = 2 \sum_{i=1}^n w_i \left\{ y_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right\},$$

where $\tilde{\theta}_i$ is the maximum likelihood estimate (MLE) of the canonical parameter corresponding to $y_i$ from the full (saturated) model; $\hat{\theta}_i$ is the MLE of the canonical parameter corresponding to $y_i$ from the reduced (proposed) Poisson GLM model, and $w_i$ is the weight from $a_i(\phi) = \phi / w_i$.

We first write the probability function of Poisson($\mu_i$) for the random variable $Y_i$ in the form of exponential dispersion family:

$$f(y_i | \mu_i) = \exp \left( y_i \log(\mu_i) - \mu_i + \log(1/y_i!) \right).$$

As a result, $a_i(\phi) = 1$, and so $w_i = 1$. $\theta_i = \log(\mu_i)$ and $b(\theta_i) = \mu_i(\theta_i) = e^{\theta_i}$.

For the full model, we know that $\tilde{\mu}_i = y_i$, and so $\tilde{\theta}_i = \log(y_i)$, $b(\tilde{\theta}_i) = \tilde{\mu}_i = y_i$. For the Poisson GLM model, $\hat{\theta}_i = \log(\hat{\mu}_i) = \log(g^{-1}(\hat{\eta}_i))$, where $\hat{\eta}_i = x_i^T \hat{\beta}$, and $b(\hat{\theta}_i) = \hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

Hence, the deviance for the Poisson GLM is

$$D(\hat{\mu}(\hat{\beta}); y) = 2 \sum_{i=1}^n \left\{ y_i (\log(y_i) - \log(\hat{\mu}_i)) - y_i + \hat{\mu}_i \right\}$$

$$= 2 \sum_{i=1}^n \left\{ y_i (\log(y_i) - \log(g^{-1}(x_i^T \hat{\beta}))) - (y_i - g^{-1}(x_i^T \hat{\beta})) \right\}$$

(b) Show that the expression for the deviance simplifies to $2 \sum_{i=1}^n y_i \log(y_i/\hat{\mu}_i)$, for the special case of the reduced model in part(a) with $g(\mu_i) = \log(\mu_i)$, and linear predictor that includes an intercept, that is, $\eta_i = \beta_1 + \sum_{j=2}^p x_{ij} \beta_j$, for $i = 1, \ldots, n$.

Solution:
If the link function is \( g(\mu_i) = \log(\mu_i) = x_i^T \beta \), \( g^{-1}(x_i^T \hat{\beta}) = \exp(x_i^T \hat{\beta}) = \hat{\mu}_i \). Now we have to show \( \sum_{i=1}^{n} (y_i - \mu_i) = \sum_{i=1}^{n} (y_i - \exp(x_i^T \hat{\beta})) = 0 \).

The log-likelihood is

\[
\ell(\beta; y) = \sum_{i=1}^{n} \{y_i x_i^T \beta - \exp(x_i^T \beta) - \log(y_i!)\},
\]

and so for all \( j = 1, \ldots, p \)

\[
\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \exp(x_i^T \beta)) x_{ij} = 0, \tag{1}
\]

If we plug \( \hat{\beta} \) into (1), we have for all \( j = 1, \ldots, p \),

\[
\sum_{i=1}^{n} (y_i - \exp(x_i^T \hat{\beta})) x_{ij} = 0.
\]

In particular, since the linear predictor includes an intercept, \( x_{i1} = 1 \) for all \( i = 1, \ldots, n \), and hence

\[
\sum_{i=1}^{n} (y_i - \exp(x_i^T \hat{\beta})) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i) = 0.
\]

Hence, the deviance is simplified to

\[
D(\hat{\mu}(\hat{\beta}); y) = 2 \sum_{i=1}^{n} \{y_i (\log(y_i) - \log(\hat{\mu}_i)) - (y_i - \hat{\mu}_i)\}
= 2 \sum_{i=1}^{n} y_i (\log(y_i/\hat{\mu}_i)) - 2 \sum_{i=1}^{n} (y_i - \hat{\mu}_i)
= 2 \sum_{i=1}^{n} y_i (\log(y_i/\hat{\mu}_i))
\]

2. Let \( y_i, i = 1, \ldots, n \), be realizations of independent random variables \( Y_i \) following \( \text{gamma}(\mu_i, \nu) \) distributions, with densities given by

\[
f(y_i | \mu_i, \nu) = \frac{(\nu/\mu_i)^\nu y_i^{\nu-1} \exp(-\nu y_i/\mu_i)}{\Gamma(\nu)}, \quad y_i > 0; \ \nu > 0, \ \mu_i > 0,
\]

where \( \Gamma(\nu) = \int_{0}^{\infty} t^{\nu-1} \exp(-t)dt \) is the Gamma function.

(a) Express the gamma distribution as a member of the exponential dispersion family.

Solution:

It is a gamma distribution with mean parameter \( \mu_i \) and shape parameter \( \nu \). The density
can be expressed as
\[
f(y_i | \nu, \mu_i) = \frac{(\nu/\mu_i)^\nu}{\Gamma(\nu)} y_i^{\nu-1} \exp\left(-\frac{(\nu/\mu_i)y_i}{1}\right)
\]
\[
= \frac{\nu^\nu}{\Gamma(\nu)} y_i^{\nu-1} \exp\left(-\frac{(1/\mu_i)y_i}{1}\right) \exp\left(\frac{\log(1/\mu_i)}{1}\right)
\]
\[
= \frac{\nu^\nu}{\Gamma(\nu)} y_i^{\nu-1} \exp\left(-\frac{(1/\mu_i)y_i - (-\log(-1/\mu_i))}{1}\right)
\]
\[
= \exp\left(-\frac{(1/\mu_i)y_i - (-\log(-1/\mu_i))}{1}\right) + \log\left(\frac{\nu^\nu}{\Gamma(\nu)} y_i^{\nu-1}\right).
\]
Let \(a_i(\phi) = \phi = 1/\nu, \theta_i = -1/\mu_i, b(\theta_i) = -\log(-1/\mu_i) = -\log(-\theta_i),\) and \(c(y_i, \phi(\nu)) = \log\left(\frac{\nu^\nu}{\Gamma(\nu)} y_i^{\nu-1}\right).\) Therefore, gamma distribution is a member of the exponential family.

(b) Obtain the scaled deviance and the deviance for the comparison of the full model, which includes a different \(\mu_i\) for each \(y_i\), with a gamma GLM based on link function \(g(\mu_i) = \eta_i = x_i^T \beta\), where \(\beta = (\beta_1, ..., \beta_p^T)\) (with \(p < n\)) is the vector of regression coefficients corresponding to a set of \(p\) covariates.

**Solution:**

First, \(a_i(\phi) = \phi = 1/\nu\) implies \(w_i = 1\) for all \(i = 1, ..., n\). \(\bar{\theta}_i = -1/\bar{\mu}_i = -1/y_i\) under the full model. Under the gamma GLM, \(\hat{\theta}_i = -1/\hat{\mu}_i\), where \(\hat{\mu}_i = g^{-1}(x_i\hat{\beta})\). Hence the deviance is

\[
D = 2 \sum_{i=1}^{n} \left\{ y_i(\hat{\theta}_i - \bar{\theta}_i) - b(\bar{\theta}_i) + b(\hat{\theta}_i) \right\}
\]
\[
= 2 \sum_{i=1}^{n} \left\{ y_i(-1/\bar{\mu}_i + 1/\hat{\mu}_i) + \log(1/\bar{\mu}_i) - \log(1/\hat{\mu}_i) \right\}
\]
\[
= 2 \sum_{i=1}^{n} \left\{ y_i(-1/y_i + 1/\hat{\mu}_i) + \log(1/y_i) - \log(1/\hat{\mu}_i) \right\}
\]
\[
= 2 \sum_{i=1}^{n} \left\{ -1 + (y_i/\hat{\mu}_i) - \log(y_i/\hat{\mu}_i) \right\}
\]
\[
= 2 \sum_{i=1}^{n} \left\{ -\log(y_i/\hat{\mu}_i) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right\}
\]

The scaled deviance is

\[
D^* = D/\phi = 2\nu \sum_{i=1}^{n} \left\{ -\log(y_i/\hat{\mu}_i) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right\}
\]

3. Consider the data set from: [http://www.stat.columbia.edu/~gelman/book/data/fabric.asc](http://www.stat.columbia.edu/~gelman/book/data/fabric.asc) on the incidence of faults in the manufacturing of rolls of fabric. The first column contains the length of each roll (the covariate with values \(x_i\)) and the second contains the number of faults (the response with means \(\mu_i\)).
(a) Use R to fit a Poisson GLM, with logarithmic link,
\[
\log(\mu_i) = \beta_1 + \beta_2 x_i
\]
(2)
to explain the number of faults in terms of length of roll.

**Solution:**

```r
##
## Call:
## glm(formula = faults ~ len_roll, family = poisson)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -2.741 -1.133 -0.039 0.662 3.075
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.971751 0.212469 4.57 4.8e-06 ***
## len_roll 0.001930 0.000306 6.30 3.0e-10 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 103.714 on 31 degrees of freedom
## Residual deviance: 61.758 on 30 degrees of freedom
## AIC: 189.1
##
## Number of Fisher Scoring iterations: 4
```

According to the output, both intercept and the covariate, length of roll, are significant. The fitted value of \( y_i \) is \( \hat{\mu}_i = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \exp(0.972 + 0.002 x_i) \). As the length of roll increases one unit, the mean number of faults will be \( e^{0.00193} \approx 1.00193 \) times larger. The residual deviance is 61.758, which is at the very right tail of \( \chi^2_{30} \) distribution, showing that the proposed Poisson GLM may be ill-fitted to the responses. For example, the mean and the variance are not equal and so an over-dispersion or under-dispersion model is more appropriate. This poor result makes the significance of \( \hat{\beta} \) questionable. The proportion of deviance explained by the model is about \( 1 - (61.758/103.714) = 0.4045 \), which can be further improved.

(b) Fit the regression model for the response means in (2) using the quasi-likelihood estimation method, which allows for a dispersion parameter in the response variance function. (Use the quasipoisson “family” in R.) Discuss the results.

**Solution:**

```r
##
## Call:
## glm(formula = faults ~ len_roll, family = quasipoisson)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -2.741 -1.133 -0.039 0.662 3.075
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.971751 0.212469 4.57 4.8e-06 ***
## len_roll 0.001930 0.000306 6.30 3.0e-10 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.001943)
##
## Null deviance: 103.714 on 31 degrees of freedom
## Residual deviance: 61.758 on 30 degrees of freedom
## AIC: 189.1
##
## Number of Fisher Scoring iterations: 4
```

According to the output, both intercept and the covariate, length of roll, are significant. The fitted value of \( y_i \) is \( \hat{\mu}_i = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \exp(0.972 + 0.002 x_i) \). As the length of roll increases one unit, the mean number of faults will be \( e^{0.00193} \approx 1.00193 \) times larger. The residual deviance is 61.758, which is at the very right tail of \( \chi^2_{30} \) distribution, showing that the proposed Poisson GLM may be ill-fitted to the responses. For example, the mean and the variance are not equal and so an over-dispersion or under-dispersion model is more appropriate. This poor result makes the significance of \( \hat{\beta} \) questionable. The proportion of deviance explained by the model is about \( 1 - (61.758/103.714) = 0.4045 \), which can be further improved.
Instead of \( \text{Var}(Y_i) = V_i(\mu) = \mu_i \) in Poisson GLM, \( \text{Var}(Y_i) = \phi V_i(\mu) \) when quasi-likelihood estimation method is used. The variance is then proportional to the mean, and the dispersion parameter \( \phi \) allows for over-dispersion if \( \phi > 1 \) and under-dispersion when \( \phi < 1 \).

The output shows that the estimated dispersion is \( \hat{\phi}_{QLE} = 2.122 > 1 \), so it is an over-dispersion model comparing to Poisson GLM. One can verify that

\[
\hat{\phi}_{QLE} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_{QLE})^2}{\hat{\mu}_{QLE}} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_{QLE})^2}{\hat{\mu}_{QLE}} = \frac{X^2}{n - p},
\]

where \( X^2 \) is generalized Pearson \( X^2 \) statistic.

Second, the point estimates for regression coefficients using quasi-likelihood method is the same as the estimates using traditional likelihood and Poisson GLM. In fact, finding quasi-likelihood estimate for \( \beta \) is independent of \( \phi \).

Notice that the standard error of \( \hat{\beta}_{QLE} \) is larger than that of \( \hat{\beta}_{MLE} \) because \( \text{Var}(\hat{\beta}_{QLE}) = \hat{\phi}_{QLE} \text{Var}(\hat{\beta}_{MLE}) \). Under over-dispersion model, the estimated coefficients have more uncertainty.

(c) Derive point estimates and asymptotic interval estimates for the linear predictor, \( \eta_0 = \beta_1 + \beta_2 x_0 \), at a new value \( x_0 \) for length of roll, under the standard (likelihood) estimation method from part (a), and the quasi-likelihood estimation method from part (b). Evaluate the point and interval estimates at \( x_0 = 500 \), and \( x_0 = 995 \). (under both cases, use the asymptotic bivariate normality of \( (\hat{\beta}_1, \hat{\beta}_2) \) to obtain the asymptotic distribution of \( \hat{\eta}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0 \).)

**Solution:**

Let \( \hat{\beta}_{MLE} = (\hat{\beta}_{1,MLE}, \hat{\beta}_{2,MLE})^T \), \( \beta = (\beta_1, \beta_2)^T \) and \( V(\hat{\beta}_{MLE}) \) be the (estimated) covariance matrix of \( \hat{\beta}_{MLE} \) under likelihood approach. Since \( \hat{\beta}_{MLE} \sim N_2\left( \beta, V(\hat{\beta}_{MLE}) \right) \)
asymptotically, \( \hat{\eta}_{0, MLE} = x_0^T \hat{\beta}_{MLE} \sim N(x_0^T \beta, x_0^T V(\hat{\beta}_{MLE}) x_0) \) asymptotically, where \( x_0 = (1, x_0)^T \).

Let \( \hat{\beta}_{QLE} = (\hat{\beta}_{1, QLE}, \hat{\beta}_{2, QLE})^T \). Notice that the covariance matrix of \( \hat{\beta}_{QLE} \) is \( \hat{\phi}_{QLE} V(\hat{\beta}_{MLE}) \), and hence asymptotically \( \hat{\eta}_{0, QLE} = x_0^T \hat{\beta}_{QLE} \sim N(x_0^T \beta, x_0^T \hat{\phi}_{QLE} V(\hat{\beta}_{MLE}) x_0) \).

The point estimates for \( \eta_0 \) are the same under likelihood and quasi-likelihood approach because both methods generate the same point estimates for \( \beta_1 \) and \( \beta_2 \). The point estimate for \( \eta_0 \) given covariate value \( x_0 \) will be

\[ \hat{\eta}_0 = \hat{\beta}_{1, MLE} + \hat{\beta}_{2, MLE} x_0 = \hat{\beta}_{1, QLE} + \hat{\beta}_{2, QLE} x_0 = 0.972 + 0.002 x_0. \]

However, the interval estimates are different under the two approaches. With asymptotical normality, the 95% confidence interval for \( \eta_0 \) is approximately

\[ \left( \hat{\eta}_{0, MLE} - 1.96 \sqrt{x_0^T V(\hat{\beta}_{MLE}) x_0}, \hat{\eta}_{0, MLE} + 1.96 \sqrt{x_0^T V(\hat{\beta}_{MLE}) x_0} \right) \]

under likelihood approach and

\[ \left( \hat{\eta}_{0, QLE} - 1.96 \sqrt{x_0^T \hat{\phi}_{QLE} V(\hat{\beta}_{MLE}) x_0}, \hat{\eta}_{0, QLE} + 1.96 \sqrt{x_0^T \hat{\phi}_{QLE} V(\hat{\beta}_{MLE}) x_0} \right) \]

under quasi-likelihood approach.

When \( x_0 = 500 \), the point estimate for \( \eta_0 \) is

\[ \hat{\eta}_{MLE} = \hat{\eta}_{QLE} = 0.972 + 0.002(500) = 1.934. \]

The corresponding 95% confidence interval is

\[ (1.934 - 1.96 \sqrt{0.006}, 1.934 + 1.96 \sqrt{0.006}) = (1.783, 2.090) \]

under likelihood method, and

\[ (1.934 - 1.96 \sqrt{2.122 \cdot 0.006}, 1.934 + 1.96 \sqrt{2.122 \cdot 0.006}) = (1.713, 2.160) \]

under quasi-likelihood method.

When \( x_0 = 995 \), the point estimate for \( \eta_0 \) is

\[ \hat{\eta}_{MLE} = \hat{\eta}_{QLE} = 0.972 + 0.002(995) = 2.892. \]

The corresponding 95% confidence interval is

\[ (2.892 - 1.96 \sqrt{0.014}, 2.892 + 1.96 \sqrt{0.014}) = (2.663, 3.121) \]

under likelihood method, and

\[ (2.892 - 1.96 \sqrt{2.122 \cdot 0.014}, 2.892 + 1.96 \sqrt{2.122 \cdot 0.014}) = (2.558, 3.226) \]

under quasi-likelihood method.
Figure 1: Scatter plots of the response and covariate $\text{fuel}$. Three different forms of response are considered: raw response, log transformation, and square root transformation. Two forms of the covariate: raw values and log transformation. Red curves are fitted cubic smoothing splines.

4. This problem deals with data collected as the number of Ceriodaphnia organisms counted in a controlled environment in which reproduction is occurring among the organisms. The experimenter places into the containers a varying concentration of a particular component of jet fuel that impairs reproduction. It is anticipated that as the concentration of jet fuel grows, the number of organisms should decrease. The problem also includes a categorical covariate introduced through use of two different strains of the organism.

The data set is available from the course website
https://ams274-fall16-01.courses.soe.ucsc.edu/node/4
where the first column includes the number of organisms, the second the concentration of jet fuel (in grams per liter), and the third the strain of the organism (with covariate values 0 and 1).

Build a Poisson GLM to study the effect of the covariates (jet fuel concentration and organism strain) on the number of Ceriodaphnia organisms. Use graphical exploratory data analysis to motivate possible choices for the link function and the linear predictor. Use classical measures of goodness-of-fit and model comparison (deviance, AIC and BIC), as well as Pearson and deviance residuals, to assess model fit and to compare different model formulations. Provide a plot of the estimation regression functions under your proposed model.

Solution:

Figure 1 shows the relationship between the response variable, the number of organisms, and the covariate, the concentration of jet fuel, under some transformations. Red curves are fitted cubic smoothing splines. One can see that, without any transformation, response and fuel have a quadratic relationship. After log or square root transformation on the response, the relationship becomes more linear. Taking log on the covariate $\text{fuel}$ makes its values separated, which is not a good idea for fitting a linear model. Hence, I focus on two cases: (1) $\log(\text{no}\_\text{organ})$ v.s. $\text{fuel}$ and (2) $\sqrt{\text{no}\_\text{organ}}$ v.s. $\text{fuel}$ for the following
analysis. That is, I use the log link function and square root link function in the Poisson GLM.

Suppose the systematic component is

$$
\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \quad i = 1, ..., n,
$$

where $\beta_1$ is the regression coefficient corresponding to the covariate fuel and $\beta_2$ is the coefficient corresponding to strain.

Here we consider two link functions: $g(\mu_i) = \log(\mu_i)$ and $g(\mu_i) = \sqrt{\mu_i}$. Table 1 shows coefficient estimates and standard errors under the two link functions. Although not shown here, all coefficients are significant with $p$-value < 0.0001. Notice that the coefficients should be interpreted in different ways.

<table>
<thead>
<tr>
<th></th>
<th>log-link Est.</th>
<th>log-link Std. Err</th>
<th>sqrt-link Est.</th>
<th>sqrt-link Std. Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>4.180</td>
<td>0.043</td>
<td>7.728</td>
<td>0.133</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-1.543</td>
<td>0.047</td>
<td>-3.590</td>
<td>0.107</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.275</td>
<td>0.048</td>
<td>0.587</td>
<td>0.120</td>
</tr>
</tbody>
</table>

Table 1: Poisson GLM coefficient estimates and standard errors under log link function and square root link function.

Take log link function for example. Given everything else, increasing one unit of concentration of jet fuel will decrease 78.628% number of organisms. Moreover, the number of organisms is 31.653% more if the strain is coded as 1. Using square root link function make it more difficult to interpret regression coefficients, but its $\hat{\beta}_1$ and $\hat{\beta}_2$ have the same sign as the estimators in log link function case. One disadvantage of using square root link function is that $\sqrt{\mu} \in \mathbb{R}^+$, but $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \in \mathbb{R}$ in general. $\sqrt{\mu}$ will become negative if the concentration of fuel is over some level, which make this model less convincing.

We now examine goodness of fit of both models. Table 2 shows (residual) deviance, AIC and BIC under the two link functions. Given that the deviance is asymptotically $\chi^2$ distributed with 67 (70 - 3) degrees of freedom, log-link is indicating less evidence of a lack of fit, while sqrt-link is more questionable. Notice that these two models cannot be compared directly because one is not nested in the other and they use different link functions.

<table>
<thead>
<tr>
<th></th>
<th>Deviance</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>log-link</td>
<td>86.376</td>
<td>415.951</td>
<td>422.696</td>
</tr>
<tr>
<td>sqrt-link</td>
<td>144.178</td>
<td>473.752</td>
<td>480.498</td>
</tr>
</tbody>
</table>

Table 2: Deviance, AIC and BIC under log link function and square root link function.

Residual plots are shown in Figure 2. Both deviance and Pearson residual plots under log link function do not show any particular pattern, but the plots under square root link function have a quadratic pattern, indicating a lack of fit.

The variance of the residuals with respect to the fitted values should also be examined. Since deviance residuals are scaled out the variance function, we use response residuals, $y - \hat{\mu}$, instead. Both log and square root link functions produce fanning effect, showing...
that variance is increasing as fitted linear predictor increases. Residuals under the square root link function still has quadratic pattern. We may need to use a quasi-likelihood GLM to explain nonconstant variance. In fact, the dispersion estimators under log link function and square root link function are 1.19 and 2.24, respectively.

Based on the analysis above, I would choose log link function for the Poisson GLM. The fitted regression plane for $\eta$ and surface for $\mu$ are shown in Figure 4.
Figure 3: Response Residual plots. Left: log link functions. Right: square root link functions.

Figure 4: Estimated regression functions. Left: Estimated linear predictor $\hat{\eta}$. Red balls indicate $\log(y_i)$. Right: Estimated mean of the response variable $\hat{\mu}$. Red balls indicate responses $y_i$. 