Homework 2

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

   (a) Obtain the expression fro the deviance for comparison of the full model, which assumes a different μ_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, $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ (with $p < n$) is the vector of regression coefficients corresponding to covariates $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^T$

   (b) Show 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 $i = 1, \dots, n$.

Sol. 1 (a) Since Y_i follows the Poisson (μ_i) distributions, let $y_i|\mu_i \overset{iid}{\sim} \frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}$. Then the log-likelihood function can be expressed as

$$\begin{aligned} \ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) &= \sum_{i=1}^n \ell(\hat{\mu}_i, y_i) && (1) \\ &= \sum_{i=1}^n e^{y_i \ln \mu_i - \mu_i - \ln y_i!} && (2) \\ & && (3) \end{aligned}$$

which shows that the nature parameter $\theta_i = \ln \mu_i$ and $b(\theta) = e^\theta$. And then $\mu_i = e^{\theta_i}$.
Next, substituting μ_i with θ_i, we have $\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) = \sum_{i=1}^n y_i\theta_i - e_i^\theta - \ln y_i!$
The deviance for Poisson GLM should be

$$\begin{aligned} D &= -2(\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) - \ell(\tilde{\boldsymbol{\theta}}; \boldsymbol{y})) && (4) \\ &= -2\sum_{i=1}^n [y_i(\hat{\theta}_i - \tilde{\theta}_i) - e^{\hat{\theta}_i} + e^{\tilde{\theta}_i}] && (5) \\ &= -2\sum_{i=1}^n [y_i \ln \frac{\hat{\mu}_i}{\tilde{\mu}_i} - \hat{\mu}_i + \tilde{\mu}_i] && (6) \end{aligned}$$

Since the full model implies that $\tilde{\mu}_i = y_i$, substituting the equation to the expression above generates that

$$D = 2\sum_{i=1}^{n}[y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)]. \tag{7}$$

(b) If $g(\mu_i) = \ln(\mu_i)$ and $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$, because $\hat{\beta}$ maximizes $\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y})$, we have $\frac{\partial \ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\beta}} = \mathbf{0}$. Because

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\beta_j} & (8)\\ &= \sum_{i=1}^{n}(y_i - e^{\theta_i})\mu_i^{-1}\mu_i x_{ij} & (9)\\ &= \sum_{i=1}^{n}(y_i - \mu_i)x_{ij} & (10) \end{aligned}$$

where $x_{i1} = 1$ for $i = 1, \ldots, n$. Because $\frac{\partial \ell}{\beta_1} = 0$, we derive that $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)x_{i1} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0$.
Therefore,

$$\begin{aligned} D &= 2\sum_{i=1}^{n}[y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)] & (11)\\ &= 2\sum_{i=1}^{n} y_i \ln \frac{y_i}{\hat{\mu}_i} & (12) \end{aligned}$$

2. (Ex. 2) Let y_i, $i = 1, \ldots, n$ be realizations of independent random variables Y_i following gamma(μ_i, ν) distributions, with densities given by

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

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.

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

Sol. 2 (a) The density function can be rewritten as

$$\begin{aligned} f(y_i|\mu_i, \nu) &= \exp(\nu \ln(\frac{\nu}{\mu_i}) + (\nu - 1)\ln y_i - \frac{\nu}{\mu_i} y_i - \ln\Gamma(\nu)) \\ &= \exp(\frac{\alpha_i y_i - \ln\frac{\alpha_i}{\theta}}{\theta} + (\frac{1}{\theta} - 1)\ln y_i - \ln\Gamma(\frac{1}{\theta})) \\ &= \exp(\frac{\alpha_i y_i - \ln\alpha_i}{\theta} + \frac{\ln\theta}{\theta} + (\frac{1}{\theta} - 1)\ln y_i - \ln\Gamma(\frac{1}{\theta})) \end{aligned}$$

where $\alpha_i = -\frac{1}{\mu_i}$ and $\theta = \frac{1}{\nu}$. It suggests that α_i is the nature parameter and θ is the disperse parameter and this distribution belongs to the exponential distribution family.

(b) The scaled deviance and deviance are derived as follows:

$$\begin{align} D^* &= -2(\ell(\hat{\boldsymbol{\mu}}, \nu; \boldsymbol{y}) - \ell(\tilde{\boldsymbol{\mu}}, \nu; \boldsymbol{y})) \tag{13} \\ &= -2\sum_{i=1}^{n}(\nu\ln(\frac{\tilde{\mu}_i}{\hat{\mu}_i}) - \frac{\nu y_i}{\hat{\mu}_i} + \frac{\nu y_i}{\tilde{\mu}_i}) \tag{14} \\ &= \nu \cdot -2\sum_{i=1}^{n}(\ln\frac{\tilde{\mu}_i}{\hat{\mu}_i} - \frac{y_i}{\hat{\mu}_i} + \frac{y_i}{\tilde{\mu}_i}) \tag{15} \\ D &= -2\sum_{i=1}^{n}(\ln\frac{\tilde{\mu}_i}{\hat{\mu}_i} - \frac{y_i}{\hat{\mu}_i} + \frac{y_i}{\tilde{\mu}_i}) \tag{16} \end{align}$$

Since $\tilde{\mu}_i = y_i$, we have that $D^* = \nu \cdot -2\sum_{i=1}^{n}(\ln\frac{y_i}{\hat{\mu}_i} - \frac{y_i}{\hat{\mu}_i} + 1)$ and $D = -2\sum_{i=1}^{n}(\ln\frac{y_i}{\hat{\mu}_i} - \frac{y_i}{\hat{\mu}_i} + 1)$, where $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$

3. (Ex. 3) Consider the data set from:
http://www.stat.columbia.edu/ gelman/book/data/fabric.asc
on the incidence of faults in the manufacturing of rolls of fabrics. 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 μ_i).

(a) Use R to fit a Poisson GLM, with logarithmic link,

$$\ln(\mu_i) = \beta_1 + \beta_2 x_i \tag{17}$$

to explain the number of faults in terms of length of roll.

Table 1: Estimation for a Poisson GLM

	estimate	std
β_0	1.04	0.21
β_1	0.00	0.00

(b) Fit the regression model for the response means in (1) 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.

(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 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 $\hat{\eta}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0$.)

(Sol. 3) (a) After using R to fit a Poisson GLM, we have the results in Table 1:

(b) The estimates of coefficients of covariates $\boldsymbol{\beta}$ are exactly same. And the disperse estimate is 2.24. This is because that the estimate of nature parameter which related to $\boldsymbol{\beta}$ is free of the disperse parameter σ. It suggests that the two methods should have the same nature parameter and so the same $\hat{\boldsymbol{\beta}}$.

(c) According to the MLE property, under the regulation condition, $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, J^{-1}(\hat{\boldsymbol{\beta}}))$. Considering the disperse parameter, $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, J^{-1}(\hat{\boldsymbol{\beta}}, \tilde{\phi}))$ where $J(\hat{\boldsymbol{\beta}}, \tilde{\phi}) = \frac{1}{\phi} J(\hat{\boldsymbol{\beta}})$.

In our case, we have $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, \begin{bmatrix} \text{4.54e-2} & \text{-6.32e-5} \\ \text{-6.32e-5} & \text{9.58e-8} \end{bmatrix})$ for no disperse parameter case and $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, \begin{bmatrix} 0.10 & \text{-1.42e-4} \\ \text{-1.42e-4} & \text{2.14e-7} \end{bmatrix})$.

Consider the distribution of $\boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ where $\boldsymbol{x} = (1, x_0)^T$, we have that

$$\boldsymbol{x}^T \hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{x}^T \boldsymbol{\beta}, \boldsymbol{x}^T \Sigma \boldsymbol{x}), \tag{18}$$

where Σ is the covariance matrix of $\hat{\boldsymbol{\beta}}$.

After computation, we conclude that:

$x_0 = 500$ (a) $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} \;\dot{\sim}\; \mathcal{N}(Y,$6.12e-3$)$ implies the 100p% confidence interval of Y should be $(\hat{Y} - \sigma * t_{\frac{1-p}{2}}, \hat{Y} + \sigma * t_{\frac{1-p}{2}}) = (1.93 - 0.08t_{\frac{1-p}{2}}, 1.93 + 0.08t_{\frac{1-p}{2}})$.

(b) $\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \;\dot{\sim}\; \mathcal{N}(Y,$1.37e-2$)$ implies the 100p% confidence interval of Y should be $(\hat{Y} - \sigma * t_{\frac{1-p}{2}}, \hat{Y} + \sigma * t_{\frac{1-p}{2}}) = (1.93 - 0.12t_{\frac{1-p}{2}}, 1.93 + 0.12t_{\frac{1-p}{2}})$.

$x_0 = 995$ (a)$\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \;\dot{\sim}\; \mathcal{N}(Y,$1.44e-2$)$ implies the 100p% confidence interval of Y should be $(\hat{Y} - \sigma * t_{\frac{1-p}{2}}, \hat{Y} + \sigma * t_{\frac{1-p}{2}}) = (2.80 - 0.12t_{\frac{1-p}{2}}, 2.80 + 0.12t_{\frac{1-p}{2}})$.

(b) $\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \;\dot{\sim}\; \mathcal{N}(Y,$3.22e-2$)$ implies the 100p% confidence interval of Y should be $(\hat{Y} - \sigma * t_{\frac{1-p}{2}}, \hat{Y} + \sigma * t_{\frac{1-p}{2}}) = (2.80 - 0.18t_{\frac{1-p}{2}}, 2.80 + 0.18t_{\frac{1-p}{2}})$.

4. (Ex. 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
   http://ams274-fall16-01.course.soe.ucsc.edu/node/4
   where the first column includes the number of organisms, the second shows 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.

Table 2: Different Poisson GLM with different link functions

	Deviance	AIC	BIC
Identity link	291.84	621.42	622.35
Log link	86.38	415.95	416.88
Inverse link	165.04	494.62	495.55
Square root link	144.18	473.75	474.68

Provide a plot of the estimated regression functions under your proposed model.

Sol. 4 Using four different link functions including identity link, log link, Inverse link and square root link, we generate the plot between the number of organisms and the concentrations of fuel for different categories in figs. 1 to 4. From those figures, we find that the log link function makes more sense, because the linear trend is more significant, which validate our model with respect to linear predictor assumption.

The deviance, AIC and BIC are summarized in the Table 2, in which the model with the identity link cannot find vailid set of coefficients in the default algorithm. Since all the deviance is asymptotically following the distribution $\chi_2 9^2$, by comparing the value of deviance, the log link function is preferable, because it is more possible to appeared in the $\chi_2 9^2$ distribution. Moreover, comparing AIC and BIC among those models, the model with log link function has the smallest value, which suggests that this model perform better than others.

Then we compute models' Pearson and deviance residuals and create plots between residuals and fitted values in figs. 5 to 12. From those residuals, it shows that the model with log link function generates the smallest residuals among all model, no matter for the Pearson residuals or deviance residuals. Hence, the residuals show that the model with log link function is the best and proposed model.

Under the proposed method, the estimated regression functions have been plotted in 13.

References

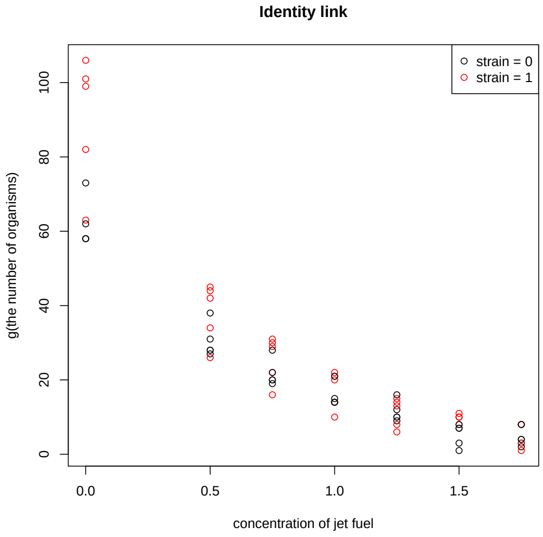

Figure 1: ”Identity link”

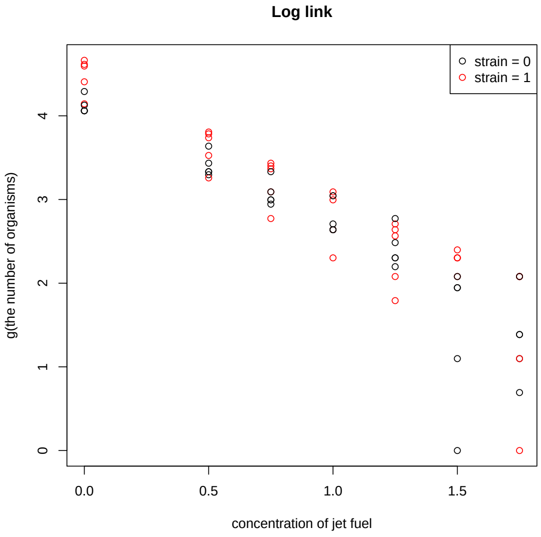

Figure 2: ”Log link”

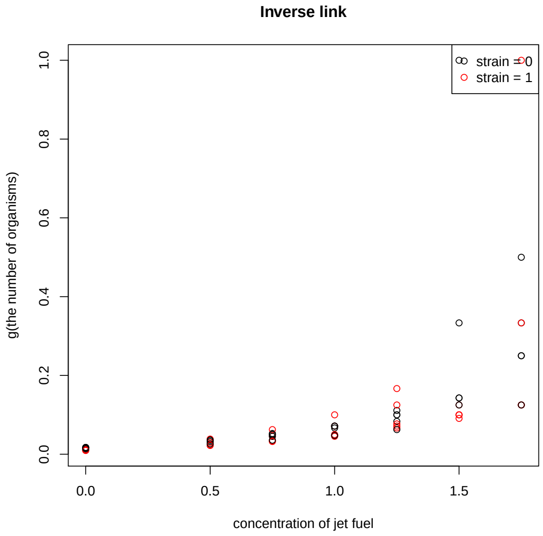

Figure 3: ”Inverse link”

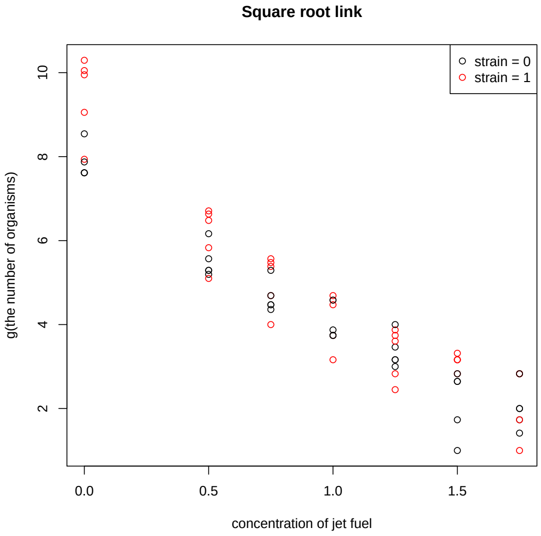

Figure 4: ”Square root link”

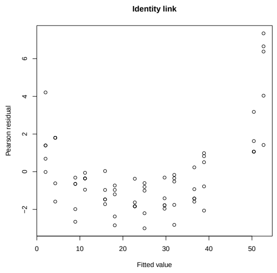

Figure 5: "Identity link"

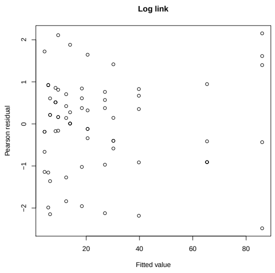

Figure 6: "Log link"

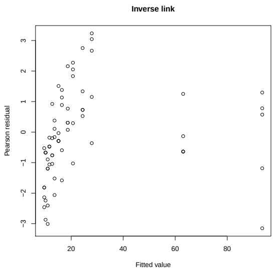

Figure 7: "Inverse link"

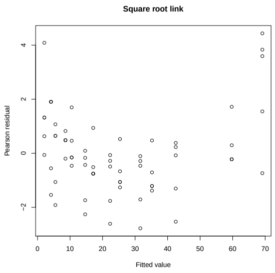

Figure 8: "Square root link"

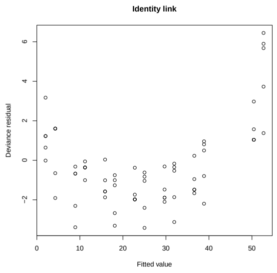

Figure 9: "Identity link"

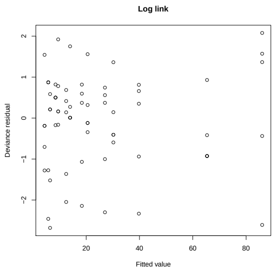

Figure 10: "Log link"

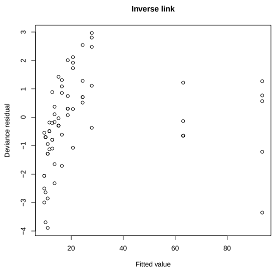

Figure 11: "Inverse link"

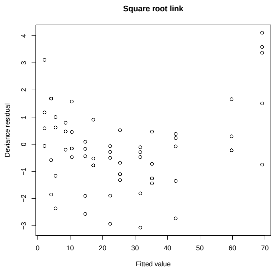

Figure 12: "Square root link"

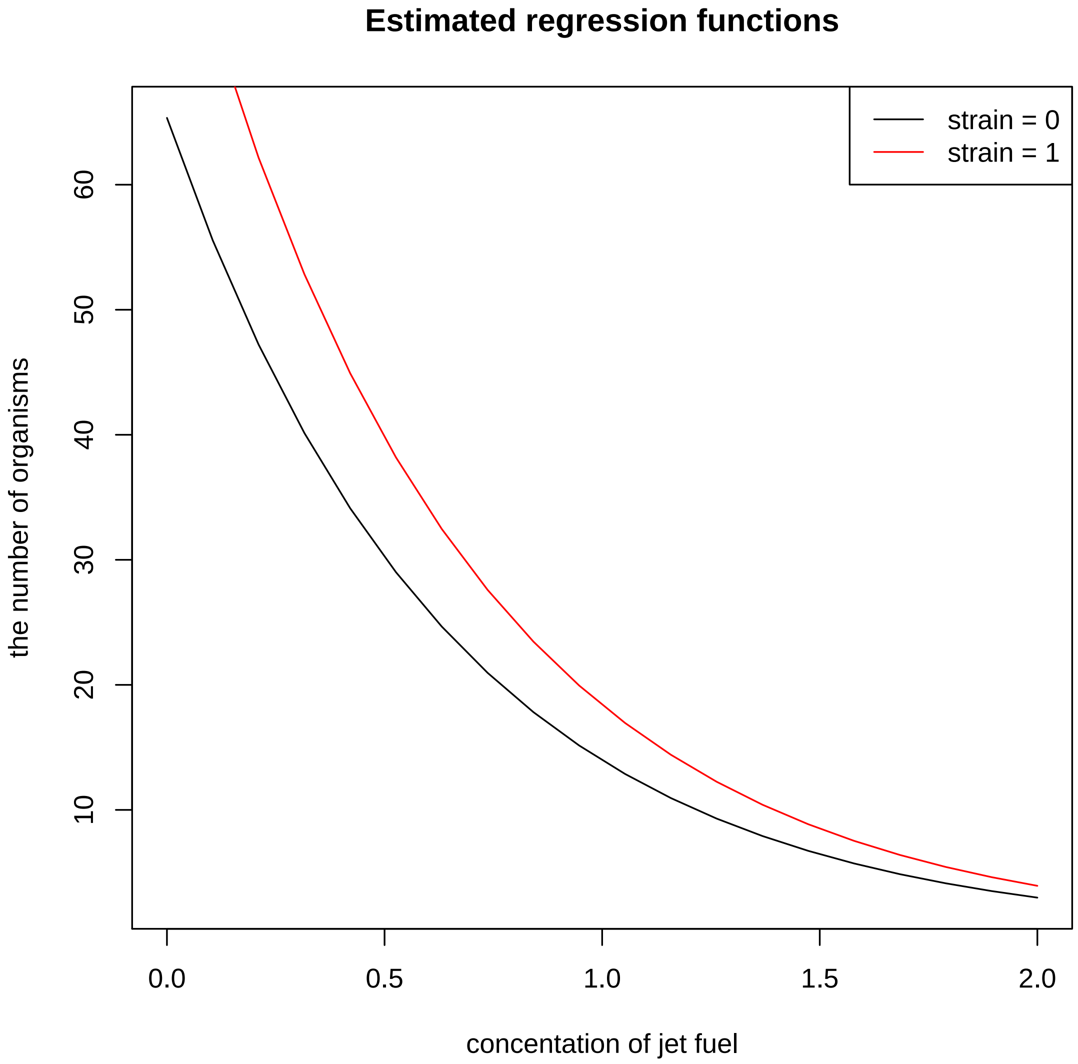

Figure 13: Estimated regression functions