1. Consider the following air pollution study with the dataset ozone.txt:

```
> data<-read.table("ozone.txt", header=TRUE, sep="\t", dec=".")
> data
    rad temp wind ozone
1   190   67  7.4    41
2   118   72  8.0    36
3   149   74 12.6    12
4   313   62 11.5    18
5   299   65  8.6    23
6    99   59 13.8    19
7    19   61 20.1     8
8   256   69  9.7    16
9   290   66  9.2    11
10  274   68 10.9    14
.
.
The dataset is gathered during the air pollution study.
The response variable is ozone. The problem is to find out,
how is ozone concentration related to wind speed, air temperature
 and intensity of solar radiation.
```

Denote the variables as following

$$Y = \text{ozone}, \quad X_1 = \text{rad}, \quad X_2 = \text{temp}, \quad X_3 = \text{wind}.$$

Note that the response variable $Y = \text{ozone}$ is continuous random variable where measurement accuracy happens to be in integer level.

(a) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the models

$$\begin{aligned}
\mathcal{M}_{\text{identity}}: &\quad \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \\
\mathcal{M}_{\text{inverse}}: &\quad \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \\
\mathcal{M}_{\log}: &\quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \\
\mathcal{M}_{\text{exponential}}: &\quad \log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) + \beta_3 \log(x_{i3}),
\end{aligned}$$

Which model fits the best to the data if the choice of model is done based on the AIC value?

i. $\mathcal{M}_{\text{identity}}$,
ii. $\mathcal{M}_{\text{inverse}}$,
iii. $\mathcal{M}_{\log}$,
iv. $\mathcal{M}_{\text{exponential}}$.

(2 points)

(b) Consider the model

$$\mathcal{M}: \quad g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Choose the link function g based on your solution to (a). Study under different distributional assumptions how Pearson's residuals are behaving. That is, consider the following linear models for Pearson residuals o_i

$$o_i^2 = \alpha_0 + \alpha_1 \hat{\mu}_i + \varepsilon_i$$

in case of normal, Gamma, and Inverse Gaussian distribution. For each distribution, test the null hypothesis $H_0 : \alpha_1 = 0$. Based on these Pearson's residuals testing results, which distributional assumption is the most suitable on?

i. $Y_i \sim N(\mu_i, \sigma^2)$,
ii. $Y_i \sim Gamma(\mu_i, \phi)$,
iii. $Y_i \sim IG(\mu_i, \phi)$.

(2 points)

(c) Regardless of your solutions to (a) and (b), let us assume $Y_i \sim Gamma(\mu_i, \phi)$ and

$$\mathcal{M}: \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Test the hypotheses

$$H_0 : \beta_2 + \beta_3 = 0,$$
$$H_1 : \beta_2 + \beta_3 \neq 0.$$

Select appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic, and return it as your answer to the question.

(2 points)

2. Consider the data set in the file weld.txt:

```
> data<-read.table("weld.txt", sep="\t", dec=".", header=TRUE)
   Drying Material Thickness Angle Opening Preheating Strength
1       0        0         0     0       0          0     43.7
2       0        1         1     1       1          1     40.2
3       1        1         0     0       0          0     42.4
4       1        0         1     1       1          1     44.7
5       1        1         0     0       1          1     42.4
6       1        0         1     1       0          0     45.9
7       0        0         0     0       1          1     42.2
8       0        1         1     1       0          0     40.6
9       1        1         0     1       0          1     42.4
10      1        0         1     0       1          0     45.5
11      0        0         0     1       0          1     43.6
12      0        1         1     0       1          0     40.6
13      0        0         0     1       1          0     44.0
14      0        1         1     0       0          1     40.2
15      1        1         0     1       1          0     42.5
16      1        0         1     0       0          1     46.5

An experiment to investigate factors affecting welding strength.

Drying - a 0-1 predictor
Material - a 0-1 predictor
Thickness - a 0-1 predictor
Angle - a 0-1 predictor
Opening - a 0-1 predictor
Preheating - a 0-1 predictor
Strength - The welding strength

G. Box and R. Meyer (1986) Dispersion effects from fractional designs, Technometrics, 28, 19-27
```

Denote variables as following

$$Y = \text{Strength},\ X_1 = \text{Drying},\ X_2 = \text{Material},\ X_3 = \text{Thickness},$$
$$X_4 = \text{Angle},\ X_5 = \text{Opening},\ X_6 = \text{Preheating}.$$

(a) First, use only variables $X_1 = \mathsf{Drying}$ and $X_2 = \mathsf{Material}$ to model the expected value of the response variable $Y = \mathsf{Strength}$. Select the appropriate default distribution for the response variable Y, and consider some competing models. After you have chosen your model, calculate the d-value for the predictive effect size difference $y_{2f} - y_{1f}$ when explanatory variables are changed from the values

$$x_{1f1} = 0, \qquad x_{1f2} = 0$$

to the values

$$x_{2f1} = 1, \qquad x_{2f2} = 1.$$

(2 points)

(b) Continue to use the same model as in (a). Since variables $X_1 = \mathsf{Drying}$ and $X_2 = \mathsf{Material}$, based on their coded values 0 and 1, are together creating 4 different subpopulations, test the hypotheses

$$H_0 : \mu_{jh} - \mu_{j_*h_*} = 0,$$
$$H_1 : \mu_{jh} - \mu_{j_*h_*} \neq 0,$$

for all possible differences

$$\mu_{00} - \mu_{01}, \mu_{00} - \mu_{10}, \mu_{00} - \mu_{11}, \mu_{10} - \mu_{01}, \mu_{10} - \mu_{11}, \mu_{01} - \mu_{11}.$$

In your answer, return the test statistic value of that pairwise comparison $\mu_{jh} - \mu_{j_*h_*}$ for which the difference is the largest.

(2 points)

(c) Consider modeling the expected value of the response variable $Y = \mathsf{Strength}$ by all the explanatory variables X_1, X_2, X_3, X_4, X_6. Select the appropriate default distribution for the response variable Y, and choose the model which you see is the most suitable one for modeling the expected value of the response variable $Y = \mathsf{Strength}$. Justify your choice of model based on one of the model selection criteria. After you have chosen your model, calculate the fitted value for the first observation, i.e., $\hat{\mu}_1$.

(2 points)

3. (a) Let us assume $Y_i \sim IG(\mu_i, \phi)$. Consider the model

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_i).$$

Let the estimates of the parameters β_0, β_1, ϕ be as $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0.5, \tilde{\phi} = 0.05$, when

$$\mathbf{X} = \begin{pmatrix} 1 & \log(3) \\ 1 & \log(3) \\ 1 & \log(3) \\ 1 & \log(6) \\ 1 & \log(6) \\ 1 & \log(6) \\ 1 & \log(9) \\ 1 & \log(9) \\ 1 & \log(9) \end{pmatrix}.$$

Calculate the estimated covariance matrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}$.

(2 points)

(b) Consider the simple Gamma model with

$$Y_i \sim Gamma(\mu_i, \phi),$$
$$\mu_i = \eta_i = \beta_0.$$

Construct the $100(1-\alpha)\%$ prediction interval for the new observation Y_f.

(2 points)

(c) Let us assume $Y_i \sim Poi(\mu_i)$. Consider the model

$$\mathcal{M}: \ \log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}.$$

Show that the deviance of the model $\mathcal{M}$ is

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

(2 points)