1. Consider the dataset in the file pigeon.txt:

```
> pigeon<-read.table(file="pigeon.txt", sep="\t", header=TRUE, dec=".")
> head(pigeon)
  PIGU Year Transect Temp  Bay     ObservCond
1    0 1986      101  4.5 Uyak           Good
2    0 1987      101  4.8 Uyak ExcellentIdeal
3    0 1988      101  4.5 Uyak ExcellentIdeal
4    0 1989      101  2.7 Uyak ExcellentIdeal
5    0 1990      101  2.2 Uyak ExcellentIdeal
6    0 1991      101  2.5 Uyak    FairAverage

Numbers of sightings of pigeon guillemot seabirds by year 1986 - 2005 at
E. Sitkalidak, Uganik, Uyak, and W. Sitkalidak by people affiliated
with the Kodiak National Wildlife Refuge, Alaska.

A data frame with 3793 observations on the following variables.
PIGU
integer count of the numbers of sightings of pigeon guillemot by transect by year
Year
integer year, 1986 - 2005
Transect
Integer code (101 - 749) for the specific plot of ground observed
Temp
a numeric vector
Bay
a factor with levels E. Sitkalidak Uganik Uyak W. Sitkalidak
ObservCond
a factor with levels ExcellentIdeal, FairAverage, and Good.
```

Denote variables as following

$$Y = \mathsf{PIGU},\ X_1 = \mathsf{Temp},\ X_2 = \mathsf{Year},\ X_3 = \mathsf{Bay},\ X_4 = \mathsf{ObservCond}.$$

Further, the column Transect identifies position where sightings have been made.

(a) Consider modeling the expected value of the response variable $Y = \mathsf{PIGU}$ by the explanatory variables X_1, X_2, X_3, X_4. Select the appropriate default distribution for the response variable Y, and consider several competing models. Choose the model which you feel is the most suitable one for modeling the expected value of the response variable $Y = \mathsf{PIGU}$. You may also consider making transformation to the original response variable Y, and then model the transformed response variable, if you find modeling of transformed variable more useful for practical point of view. The best model you choose does not have to include all possible explanatory variables. In your solution, try to report your modeling process, that is, which models you considered, which link functions you compared, which goodness of fit measures you used, and which hypotheses you tested before you chose your final model. Reporting can mainly be R-code with some clarifying comments. Try make sure that your R-code is running without errors before returning your solution.

(3 points)

(b) After you have chosen your model, calculate the maximum likelihood estimate for the new observation y_f when

Year	Transect	Temp	Bay	ObservCond
2005	726	4.5	W. Sitkalidak	ExcellentIdeal

Left out those values of explanatory variables which are not included in your model. **Create also some useful prediction intervals for the new observation** y_f.

(2 points)

(c) Test at 5% significance level, is the explanatory variable $X_4 = \mathsf{ObservCond}$ statistically significant variable. In your solution, report clearly how you calculated the value of the test statistic.

(1 point)

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

```
> pigeon<-read.table(file="pigeon.txt", sep="\t", header=TRUE, dec=".")
> head(pigeon)
  PIGU Year Transect Temp  Bay     ObservCond
1    0 1986      101  4.5 Uyak           Good
2    0 1987      101  4.8 Uyak ExcellentIdeal
3    0 1988      101  4.5 Uyak ExcellentIdeal
4    0 1989      101  2.7 Uyak ExcellentIdeal
5    0 1990      101  2.2 Uyak ExcellentIdeal
6    0 1991      101  2.5 Uyak    FairAverage


source: A.N. Garand and L.N. Bell (1997). "Caffeine Content of Fountain and
Private-Label Store Brand Carbonated Beverages," Journal of the
American Dietetic Association, Vol. 97, #2, pp. 179-182.

Description: Caffeine content (mg/12oz) for 2 formulations (sugar/diet)
of 2 Brands (Coca-Cola/Pepsi) from 12 restaurants (5 Coca-Cola, 7 Pepsi).
10 replicates per restaurant per formulation. Note that restaurants are
nested within brand, not formulation.

Variables/Columns:
Brand    1=Coke, 2=Pepsi
Formulation   1=Sugar,  2=Diet
Restaurant    1=Red Lobster, 2=Applebees, 3=McDs, 4=BK, 5=Hardees
  6=Arbys, 7=Subway2, 8=Subway1,  9=KFC, 10=PizzaHut, 11=TacoBell, 12=Wendys
Caffeine Content  (mg/12oz)   34-40
```

Denote variables as following

$$Y = \text{Caffeine},\ X_1 = \text{Brand},\ X_2 = \text{Formulation},\ X_3 = \text{Restaurant}.$$

(a) Consider modeling the expected value of the response variable $Y = \mathsf{Caffeine}$ by the explanatory variables X_1, X_2, X_3. Select the appropriate default distribution for the response variable Y, and consider several competing models. Choose the model which you feel is the most suitable one for modeling the expected value of the response variable $Y = \mathsf{Caffeine}$. You may also consider making transformation to the original response variable Y, and then model the transformed response variable, if you find modeling of transformed variable more useful for practical point of view. The best model you choose does not have to include all possible explanatory variables. In your solution, try to report your modeling process, that is, which models you considered, which link functions you compared, which goodness of fit measures you used, and which hypotheses you tested before you chose your final model. Reporting can mainly be R-code with some clarifying comments. Try make sure that your R-code is running without errors before returning your solution.

(3 points)

(b) After you have chosen your model, construct 80% prediction interval for the (predictive) effect size difference $y_{2f} - y_{1f}$ when explanatory variables are changed from the values

$$x_{1f1} = \text{Coke}, \qquad x_{1f2} = \text{Diet}$$

to the values

$$x_{2f1} = \text{Pepsi}, \qquad x_{2f2} = \text{Diet}.$$

When considering predictive effect size, you can choose the value of variable $X_3 = \mathsf{Restaurant}$ any level you want yourself.

(2 points)

(c) Test at 5% significance level, is the explanatory variable $X_1 = \mathsf{Brand}$ statistically significant variable. In your solution, report clearly how you calculated the value of the test statistic.

(1 point)