4. The following data were collected on the number of aluminum cans damaged during shipping on a semitruck and the distance shipped, in hundreds of miles.

Distance (x_j)	4	3	5	8	4	3	3	4	3	5	7	3	8
Cans (Y_j)	27	54	86	136	65	109	28	75	53	33	168	47	52

Let $Y_1, Y_2, ..., Y_n$ denote independent Poisson random variables, such that Y_j has mean $\lambda_j > 0$, where Y_j = the number of cans damaged during shipment j. Consider modeling the relationship between the mean number of damaged cans, λ_j, and the distance of the shipment, x_j, as

$$\log(\lambda_j) = \alpha + \beta x_j$$

where $x_1, ..., x_n$ are assumed to be known constants and α and β are unknown parameters.

(a) Sketch a picture of Y versus x on a scatterplot. Comment on the underlying relationship between Y and x. On the scatterplot, sketch what you think $\mu_{Y|x} = E[Y|x]$ is in terms of prediction.

(b) Explain why a log transformation should make the conditional mean more linear.

(c) Determine the likelihood function, $L(\boldsymbol{\lambda}) = L(\lambda_1, ..., \lambda_n)$.

(d) Determine the log-likelihood function $l(\boldsymbol{\lambda}) = l(\lambda_1, ..., \lambda_n)$.

(e) Substitute $\log(\lambda_j) = \alpha + \beta x_j$ into the log-likelihood function to determine the log-likelihood function $l(\alpha, \beta)$.

(f) Determine the 2 non-linear functions that need to be solved numerically to determine the maximum likelihood estimates (MLEs) of α and β.

(g) Use the `R` code given on the next page (and provided on the exam website in the file *cans.R*) to determine the values of the MLEs.

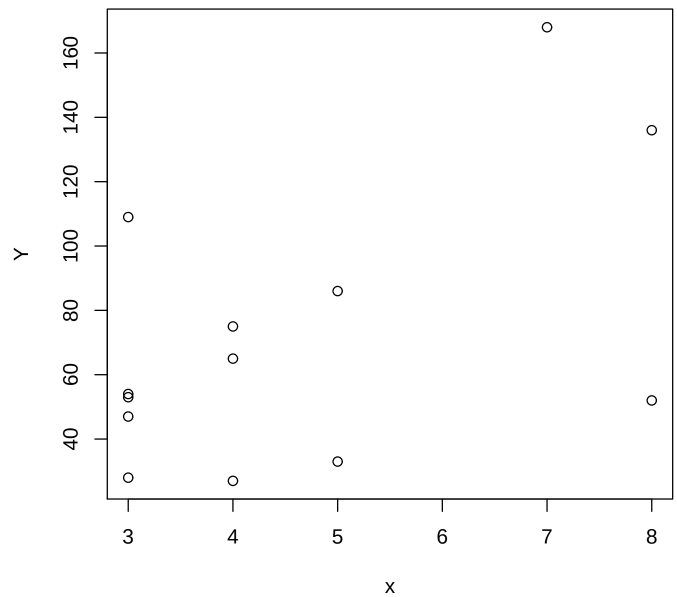
(h) Conduct a generalized likelihood ratio test for $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

```
# Check to see if these libraries are installed in R.
# If not, run the next two lines of code to install them or use the
# pull-down menu Packages > Install package(s)... .

library(ismev)
library(stats4)

# install.packages("ismev", repos = "http://cran.cnr.Berkeley.edu")
# install.packages("stats4", repos = "http://cran.cnr.Berkeley.edu")

x = c(4, 3, 5, 8, 4, 3, 3, 4, 3, 5, 7, 3, 8)
Y = c(27, 54, 86, 136, 65, 109, 28, 75, 53, 33, 168, 47, 52)

n = length(Y);  n

X11(); plot(x, Y)

Y.log = log(Y)

X11(); plot(x,Y.log)

# minus the log likelihood

ll = function(a,b){
  -sum(Y*(a + b*x)) + sum(exp(a + b*x)) + sum(log(factorial(Y)))
}

model.mle = mle(minuslog=ll,start=list(a=1,b=1));  model.mle

# plot fitted model

a.mle = coef(model.mle)[1];  a.mle
b.mle = coef(model.mle)[2];  b.mle

x.index = seq(min(x), max(x),0.01)
Y.fit = exp(a.mle + b.mle*x.index)

plot(x,Y,xlab="speed",ylab="damaged cans",main="Fitted Model")
lines(x.index,Y.fit,type="l",col=3)

a.0 = log(mean(Y));  a.0

LR.stat = 2*n*(a.mle - a.0)*mean(Y) + 2*b.mle*sum(x*Y)
LR.stat

qchisq(.95,1)
```

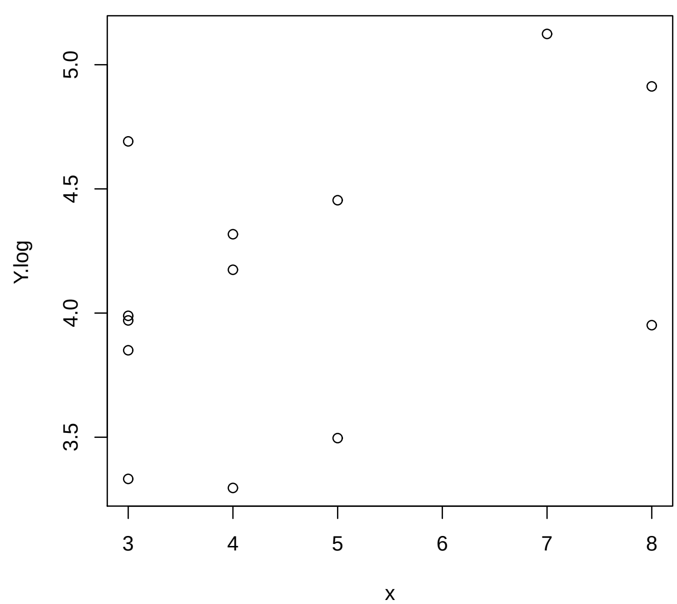
4. (a) The plot seems to have an increasing trend, that is curved upward, and it shows heteroskedasticity.

There seems to be an outlier at $(3, 109)$.

So the best predictor $\mu_{Y|x}$ should be curved upward.



(b) A log transformation should help to improve the linearity of the plot $log(Y_j)$ versus x_j.



(c) Let $Y_j \sim Poisson(\lambda_j)$, where $\lambda_j > 0$ and $j = 1, ..., n$.

$$f(y_j) = \lambda_j^{y_j} \frac{e^{-\lambda_j}}{y_j!}$$

So

$$\begin{aligned} L(\boldsymbol{\lambda}) &= L(\lambda_1, ..., \lambda_n) \\ &= \prod_{j=1}^{n} f(y_j) \\ &= \prod_{j=1}^{n} \lambda_j^{y_j} \frac{e^{-\lambda_j}}{y_j!} \end{aligned}$$

(d)

$$\begin{aligned} l(\boldsymbol{\lambda}) &= l(\lambda_1, ..., \lambda_n) \\ &= \sum_{j=1}^{n} y_j \log(\lambda_j) - \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{n} \log(y_j!) \end{aligned}$$

(e)

$$l(\alpha, \beta) = \sum_{j=1}^{n} y_j(\alpha + \beta x_j) - \sum_{j=1}^{n} \exp(\alpha + \beta x_j) - \sum_{j=1}^{n} \log(y_j!)$$

(f) Taking partial derivatives with respect to α and β and setting each equation equal to zero, yields the following two equations.

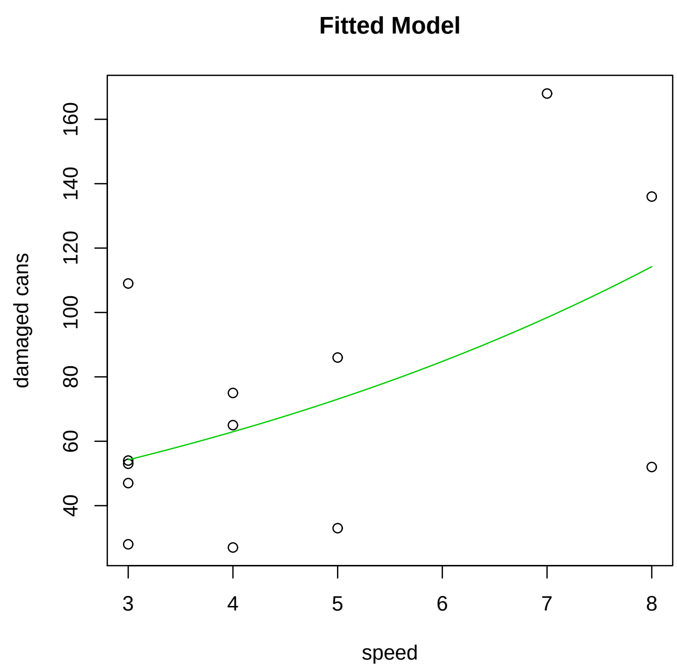
$$\sum_{j=1}^{n} \exp(\alpha + \beta x_j) = \sum_{j=1}^{n} y_j$$

$$\sum_{j=1}^{n} x_j \exp(\alpha + \beta x_j) = \sum_{j=1}^{n} x_j y_j$$

(g) Using the `mle()` function in, provided in the `R` code, $\hat{\alpha} = 3.546$ and $\hat{\beta} = 0.149$.

```
> # minus the log likelihood
>
> ll = function(a,b){
+   -sum(Y*(a + b*x)) + sum(exp(a + b*x)) + sum(log(factorial(Y)))
+ }

> model.mle = mle(minuslog=ll,start=list(a=1,b=1))
> model.mle

Call:
mle(minuslogl = ll, start = list(a = 1, b = 1))

Coefficients:
        a         b
3.5464014 0.1489722
```

The plot of the fitted model.



(h) Test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.
The GLR in general

$$\Lambda = \frac{\max_{\Omega_0} L(\theta)}{\max_{\Omega} L(\theta)}$$

here

$$
\begin{aligned}
\Lambda &= \frac{\max_{\alpha_0} L(\alpha_0)}{\max_{\alpha,\beta} L(\alpha,\beta)} \\
&= \frac{L(\hat{\alpha}_0)}{L(\hat{\alpha},\hat{\beta})}
\end{aligned}
$$

Under the null hypothesis, $\hat{\alpha}_0 = \log(\bar{y}) = 4.27$. Under the alternative hypothesis the MLEs were computed above, $\hat{\alpha} = 3.546$ and $\hat{\beta} = .149$.
The LR statistic is

$$
-2\log(\Lambda) = 2n(\hat{\alpha} - \hat{\alpha}_0)\bar{y} + 2\hat{\beta}\sum x_j y_j
$$

and from the `R` output, the computed value of the LR statistics is 78.22.

```
> LR.stat = 2*n*(a.mle - a.0)*mean(Y) + 2*b.mle*sum(x*Y)
> LR.stat

78.21663
```

Recall that the LR statistic $-2\log(\Lambda)$ has a Chi-Square distribution, here with degrees of freedom equal to 1. So to conduct the hypothesis test we need the critical value from this distribution with a significance level $\alpha = 0.05$.

```
> qchisq(.95,1)
[1] 3.841459
```

Reject H_0 at the 5% significance level, because the likelihood ratio statistics (78.22) is greater than the Chisquare, df = 1, critical value (3.84).
There is evidence that the number of damaged can's is related to the speed of the semitruck.