# Chapter Four

*Ryan Honea*

## Exercise One

### Question

The data in Table the table below show the numbers of cases of AIDS in Australia by date of diagnosis for successive 3-months periods from 1984 to 1988. (Data from National Centre for HIV Epidemiology and Clinical Research 1994.)

In this early phase of the epidemic, the numbers of cases seemed to be increasing exponentially.

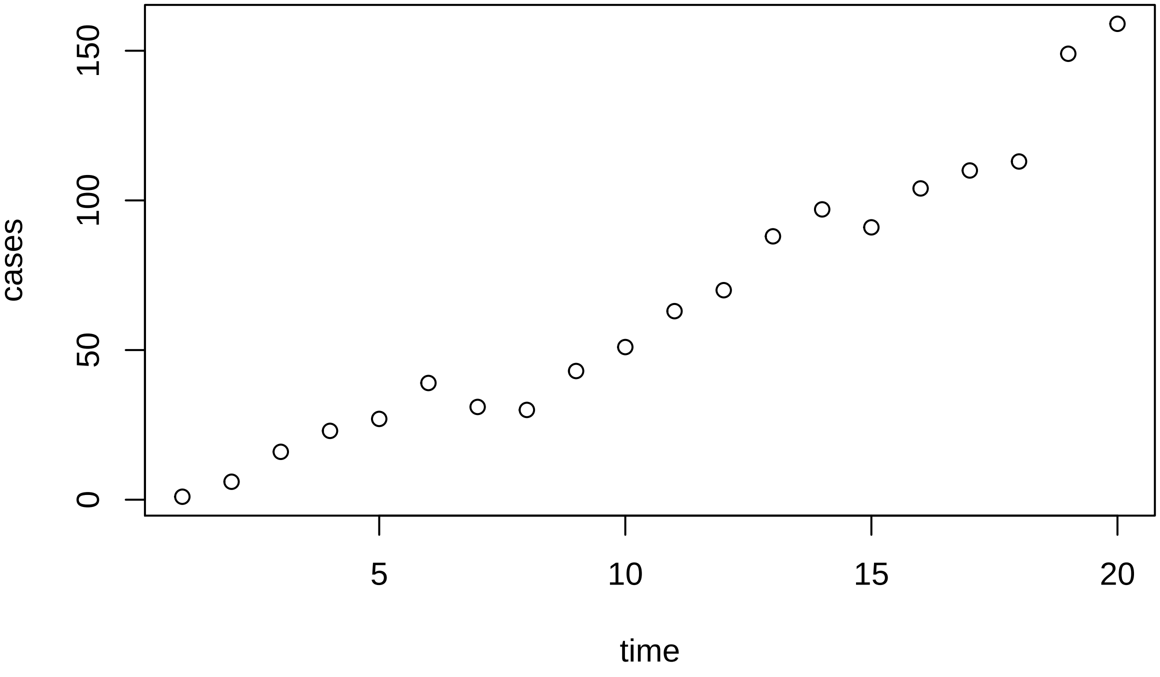
| | Quarter | | | |
|---|---|---|---|---|
| Year | 1 | 2 | 3 | 4 |
| 1984 | 1 | 6 | 16 | 23 |
| 1985 | 27 | 39 | 31 | 30 |
| 1986 | 43 | 51 | 63 | 70 |
| 1987 | 88 | 97 | 91 | 104 |
| 1988 | 110 | 113 | 149 | 159 |

### Solutions

**(a)**: Plot the number of cases $y_i$ against time period $i$ ($i = 1, ..., 20$).

*Solution:*

```
time <- seq(1,20,by=1)
cases <- c(01, 06, 16, 23, 27, 039, 031, 030, 043, 051,
          63, 70, 88, 97, 91, 104, 110, 113, 149, 159)
plot(cases ~ time)
```



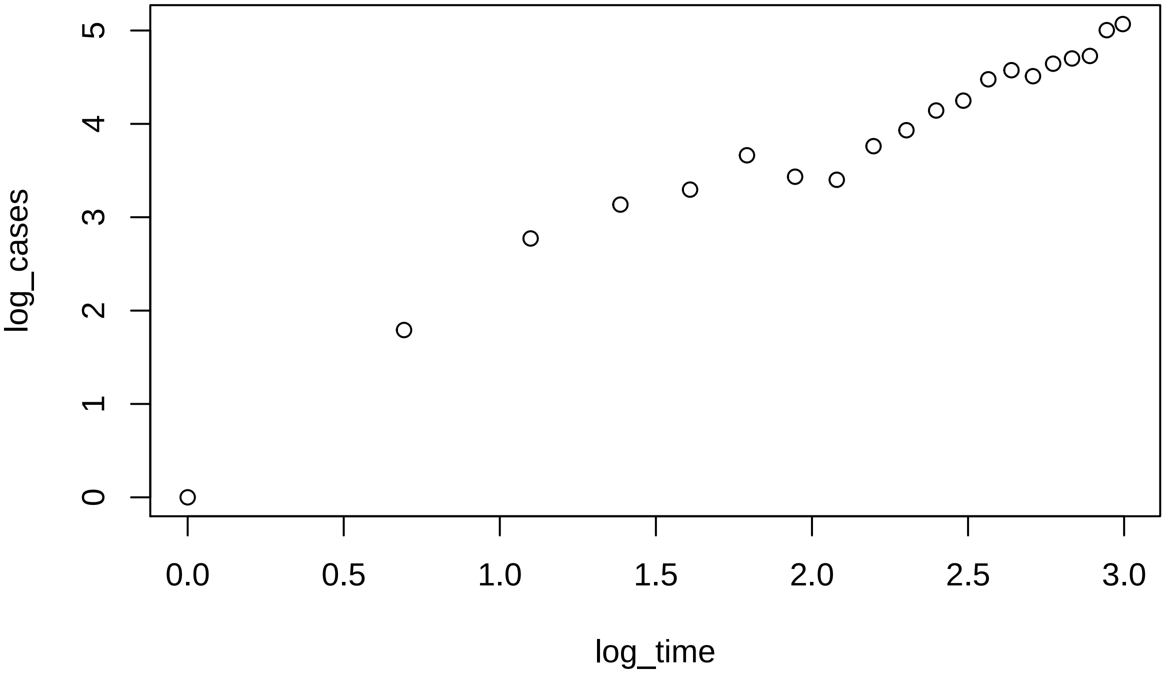
**(b)**: A possible model is the Poisson distribution with parameter $\lambda_i = i^\theta$, or equivelantly

$$\log \lambda_i = \theta \log i.$$

Plot $\log y_i$ against $\log i$ to examine this model.

*Solution:*

```
log_cases = log(cases)
log_time = log(time)
plot(log_time, log_cases)
```



**(c)**: Fit a generalized linear model to these data using the Poisson distribution, the log-link function and the equation

$$g(\lambda_i) = \log \lambda_i = \beta_1 + \beta_2 x_i,$$

where $x_i = \log i$. Firstly, do this from first principles, working out expressions for the weight matrix $\boldsymbol{W}$ and other terms needed for the iterative equation

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{b}^{(m)} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$$

and using software which can perform matrix operations to carry out the calculations.

*Solution:* Note that

$$w_{ii} = \frac{1}{\mathrm{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

and

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)$$

and these are the values for which we must find solutions.

In this case, $\mu_i = \lambda_i = e^{\beta_1 + \beta_2 x_i} = e^{\eta_i}$ and thus $\frac{\partial \mu_i}{\partial \eta_i} = e^{\eta_i}$. We also know that $\mathrm{var}(Y_i)$ of a Poisson is just $\lambda_i$ (where $\lambda_i = e^{\eta_i}$ here). Therefore we have

$$w_{ii} = \frac{1}{e^{\eta_i}} \left( e^{\eta_i} \right)^2 = e^{\eta_i} = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}.$$

and

$$\begin{aligned} z_i &= \boldsymbol{x}_i\boldsymbol{\beta} + (y_i - e^{\eta_i})(e^{\eta_i}) \\ &= \eta_i + (y_i - e^{\eta_i})e^{-\eta_i} \\ &= \eta_i + \frac{y_i}{\eta_i} - 1 \end{aligned}$$

We can therefore solve this with the following iterative R code initializing our $\boldsymbol{\beta}$ to the means of the $x_i$.

```
require(MASS)
```

```
## Loading required package: MASS
```

```
x = as.matrix(log(time))
y = as.matrix(cases)
x = cbind(rep(1,length(x)), x)
b = as.matrix(c(mean(x[,1]),mean(x[,2])))
for (i in 1:1000) {
  oldb = b
  w_ii = exp(x%*%b)
  W = sqrt(w_ii) %*% t(sqrt(w_ii)) * diag(length(x[,1]))
  z = x %*% b + y/exp(x%*%b) - 1
  b = ginv(t(x) %*% W %*% x) %*% (t(x) %*% W %*% z)
  b1 = sprintf("%.5f", b[1])
  b2 = sprintf("%.5f", b[2])
  cat("Iteration", i, ": b1 = ", b1, ", b2 = ", b2, "\n")
  if (abs(oldb[1] - b[1]) < .0005 & abs(oldb[2] - b[2]) < .0005) {
    cat("Converged at", i, "iterations")
    break
  }
}
```

```
## Iteration 1 : b1 =  0.47222 , b2 =  1.98729
## Iteration 2 : b1 =  0.43624 , b2 =  1.73805
## Iteration 3 : b1 =  0.79247 , b2 =  1.45112
## Iteration 4 : b1 =  0.98230 , b2 =  1.33503
## Iteration 5 : b1 =  0.99594 , b2 =  1.32665
## Iteration 6 : b1 =  0.99600 , b2 =  1.32661
## Converged at 6 iterations
```

**(d)**: Fit the model in (c) using statistical software which can perform Poisson regression. Compare the results with those obtained in (c).

```
glmod <- glm(cases ~ log(time), family = "poisson")
cat("b1 =", glmod$coefficients[1], "\nb2 =", glmod$coefficients[2])
```

```
## b1 = 0.995998
## b2 = 1.32661
```

The results of the model using the built in `glm` feature in R are the same as the results from the code in part **c**.

## Exercise Two

### Question

The data in the table below are times to death, $y_i$, in weeks from diagnosis and $\log_{10}$(initial white blood cell count), $x_i$, for seventeen patients suffering from leukemia. (This is Example U from Cox and Snell 1981.)

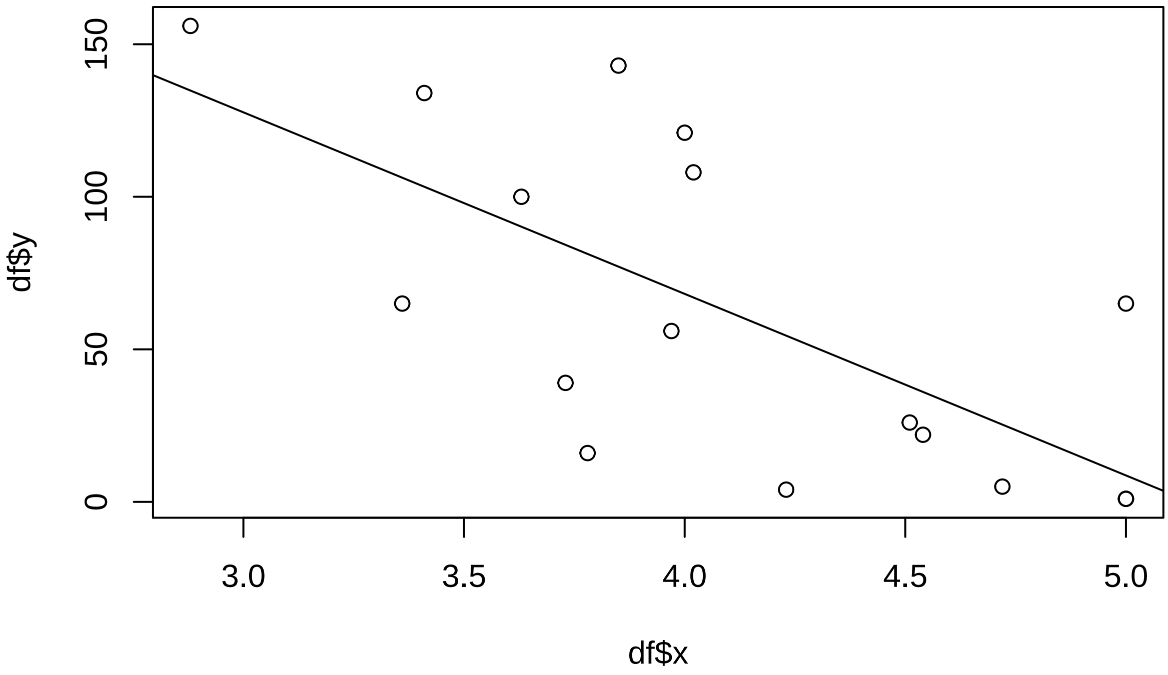
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 65 | 156 | 100 | 134 | 16 | 108 | 121 | 4 | 39 |
| $x_i$ | 3.36 | 2.88 | 3.63 | 3.41 | 3.78 | 4.02 | 4.00 | 4.23 | 3.73 |
| $y_i$ | 143 | 56 | 26 | 22 | 1 | 1 | 5 | 65 | |
| $x_i$ | 3.85 | 3.97 | 4.51 | 4.54 | 5.00 | 5.00 | 4.72 | 5.00 | |

### Solutions

**(a)**: Plot $y_i$ against $x_i$. Do the data show any trends?

*Solution:*

```r
x <- c(3.36, 2.88, 3.63, 3.41, 3.78, 4.02, 4.00, 4.23, 3.73,
       3.85, 3.97, 4.51, 4.54, 5.00, 5.00, 4.72, 5.00)
y <- c(065, 156, 100, 134, 016, 108, 121, 004, 039,
       143, 056, 026, 022, 001, 001, 005, 065)
df <- data.frame(cbind(x, y))
plot(df$x, df$y)
abline(lm(df$y ~ df$x))
```



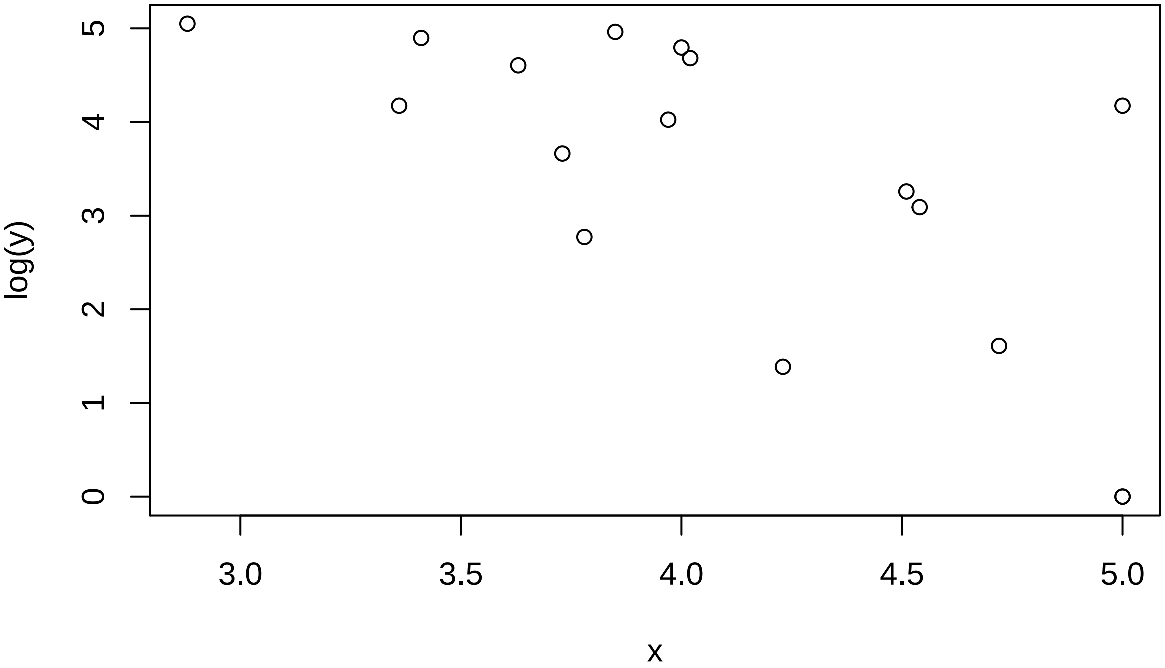
I decided to go ahead and add the line that would come from a typical linear regression. The only clear trend seems to be a slightly downward trend, but overall the data seems to be pretty dispersed as it stands right now.

**(b)**: A possible specification for $E(Y)$ is

$$\mathrm{E}(Y_i) = \exp(\beta_1 + \beta_2 x_i),$$

which can ensure all $\mathrm{E}(Y)$ is non-negative for all values of the paramaters and all values of $x$. Which link function is appropriate in this case?

```
plot(x, log(y))
```



This kind of helps in visualizing a potential model to describe our results. Again, it seems that the log-link might be our best result.

**(c)**: The Exponential distribution is often used to describe survival times. The probability distribution is $f(y;\theta) = \theta e^{-y\theta}$. This is a sepcial case of the Gamma distribution with shape parameters $\phi = 1$ (see Exercise 3.12(a)). Show that $\mathrm{E}(Y) = 1/\theta$ and $\mathrm{var}(Y) = 1/\theta^2$.

*Solution:* In Exercise 3.12(a), it was shown that the Exponential is a special case of the Gamma distribution where $\alpha = 1$, and $\beta = \theta$. The expected value and mean of the gamma distribution are $\mathrm{E}[Y] = \frac{\alpha}{\theta}$ and $\mathrm{var}[Y] = \frac{\alpha}{\beta^2}$. Therfore, $\mathrm{E}[Y] = \frac{1}{\theta}$ and $\mathrm{var}[Y] = \frac{1}{\theta^2}$.

**(d)**: Fit a model with the equation for $\mathrm{E}(Y_i)$ given in (b) and the Exponential distribution using appropriate statistical software.

```
glmod <- glm(y ~ x, family = Gamma(link = "log"))
cat("b1 =", glmod$coefficients[1], "\nb2 =", glmod$coefficients[2])
```

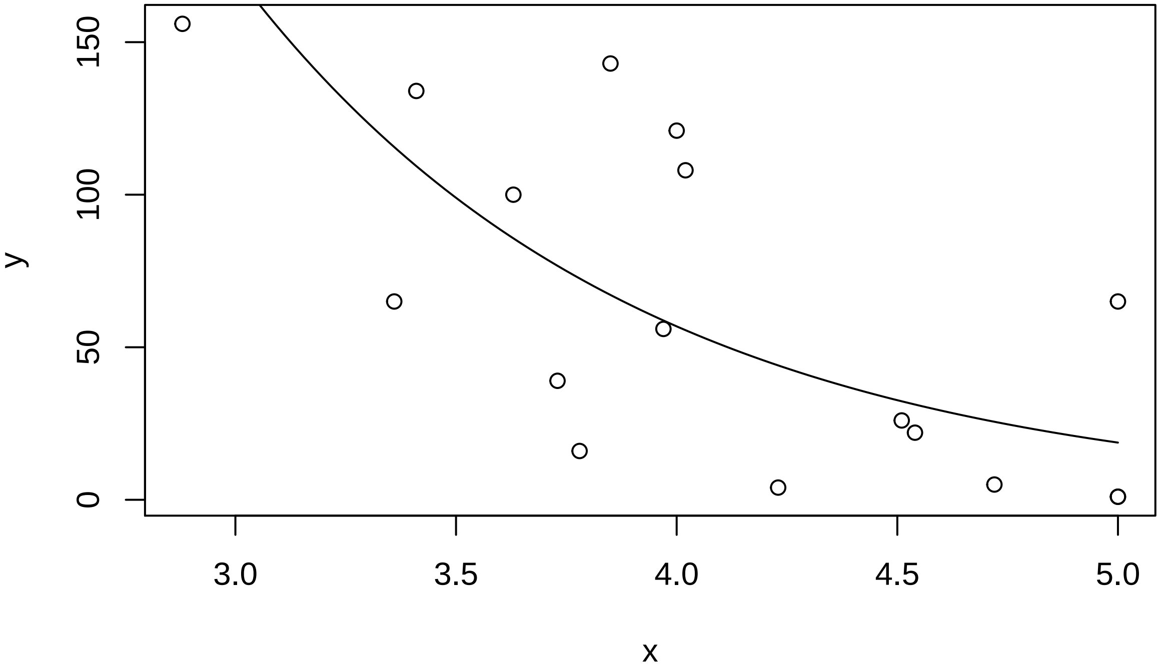
```
## b1 = 8.477494
## b2 = -1.109297
```

**(e)**: For the model fitted in (d), compare the observed values $y_i$, and fitted values $\hat{y}_i = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)$, and use the standardized residuals $r_i = (y_i - \hat{y}_i)/\hat{y}_i$ to investigate the adequacy of the model. (Note: $\hat{y}_i$ is used as the denominator of $r_i$ because it is an estimate of the standard deviation of $Y_i$–see (c) above.)

```
predictions = predict(glmod, list(x = x), type= "resp")
residuals = (y - predictions)/predictions
residuals
```

```
##           1           2           3           4           5           6
## -0.43778372 -0.20773747  0.16698607  0.22513214 -0.77947903  0.94256889
##           7           8           9          10          11          12
##  1.12864291 -0.90917977 -0.49148188  1.13004731 -0.04708838 -0.19464163
##          13          14          15          16          17
## -0.29548320 -0.94665681 -0.94665681 -0.80449595  2.46730734
```

```
plot(x,y)
curve(predict(glmod, list(x = x), type= "resp"), add=TRUE)
```



Based on these standardized residuals which are not too far out of reason, it appears that this fit is actually quite adequate. The plot seems to show that there are issues, but considering the high variance of the data, this is perhaps the most adequate fit that is not also in danger of overfitting.