8.3 (Calculation of LDFs for correlated x_1 and x_2) The inverse of $\boldsymbol{S}$ equals

$$\boldsymbol{S}^{-1} = \begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}.$$

So the LDFs are

$$\begin{aligned} L_1 &= (7,6)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(7,6)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} 7 \\ 6 \end{bmatrix} \\ &= \frac{5}{3}x_1 + \frac{1}{9}x_2 - \frac{37}{6} \end{aligned}$$

and

$$\begin{aligned} L_2 &= (4,3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(4,3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix} \\ &= x_1 - 2. \end{aligned}$$

For a new observation $\boldsymbol{x} = (x_1, x_2)' = (5,5)'$, we have $L_1 = (5/3)5 + (1/9)5 - (37/6) = 2.722$ and $L_2 = 5 - 2 = 3.000$. Since $L_1 < L_2$, this observation is classified to group 2.

LDs are given by

$$\text{LD} = (7-4, 6-3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

For the new observation $\boldsymbol{x} = (x_1, x_2)' = (5,5)'$, we have $\text{LD} = (2/3)5 + (1/9)5 = 3.889$. Furthermore, $\text{LD}_1 = (2/3)7 + (1/9)6 = 5.333$ and $\text{LD}_2 = (2/3)4 + (1/9)3 = 3.000$, so $\overline{\text{LD}} = 4.167$. Since $\text{LD} < \overline{\text{LD}}$ the observation is classified to group 2.

8.4 (Coronary heart disease data):

(a) Because the coefficients of the LDF for CHD are > those for the LDF for NCHD.

(b)

$L_{\text{NCHD}} = -23.561 + 0.027 \times 50 + 0.338 \times 95 + 0.075 \times 210 = 25.649$
and
$L_{\text{CHD}} = -238.726 + 0.072 \times 50 + 0.360 \times 95 + 0.079 \times 210 = 25.664.$
Since $L_{\text{CHD}} > L_{\text{NCHD}}$, classify the observation to the CHD group.

(c)

$$\widehat{p}_{\text{CHD}} = \frac{\exp(25.664)}{\exp(25.664) + \exp(25.649)} = 0.5038.$$

Hence

$$\widehat{p}_{\text{NCHD}} = 1 - 0.5038 = 0.4962.$$

Since $\widehat{p}_{\text{CHD}} > \widehat{p}_{\text{NCHD}}$, classify the observation to the CHD group.

(d) The Euclidean distances are
$d_{\text{NCHD}} = [(50-44.81)^2 + (95-86.99)^2 + (210-201.27)^2]^{1/2} = 12.935,$

and
$d_{\text{CHD}} = [(50-56.86)^2+(95-95.62)^2+(210-221.51)^2]^{1/2} = 13.414.$
Since $d_{\text{NCHD}} < d_{\text{CHD}}$, classify the observation to the NCHD group. You get the opposite result because the Euclidean distance does not take into account the covariance matrix as does the Mahalnobis distance.

8.5 (Fisher's iris data): The R program for this exercise is given below.

```
library(MASS)
Iris=read.csv("c:/data/Iris.csv")
fit=lda(Species_No~Petal_width+Petal_length+Sepal_length
    +Sepal_width, data=Iris, prior=c(1,1,1)/3)
fit
predict(fit,newdata=data.frame(Petal_width=1.5,Petal_length=4.0,
       Sepal_length=5.5,
Sepal_width=3.0))
```

(a) Fisher's LDFs are given in the following output.

```
Coefficients of linear discriminants:
                    LD1         LD2
Petal_width  -2.8104603 -2.83918785
Petal_length -2.2012117  0.93192121
Sepal_length  0.8293776 -0.02410215
Sepal_width   1.5344731 -2.16452123

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

Thus the first LDF achieves more than 99% of the discrimination.

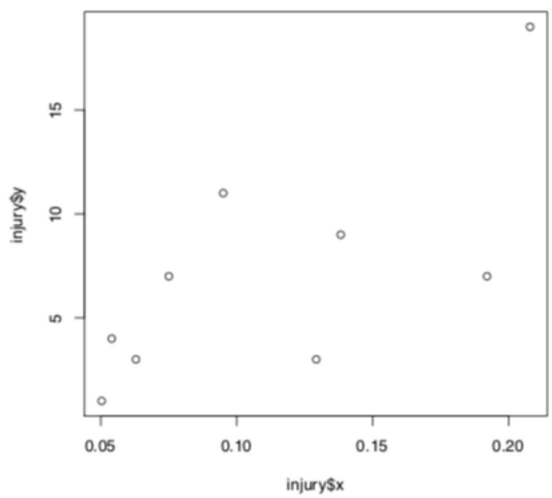
(b) The `predict` function gives the following posterior probabilities for the three species of irises:

```
$posterior
             1         2            3
1 1.962381e-19 0.9993753 0.0006246577
```

Thus in all likelihood (> 0.999) this iris belongs to species $\sharp 2$: iris virginica.

9.3 (Airline injury incidents):

(a) As the proportion of flights increases, the number of injuries increases. The variance seems to increase with the number of injuries.

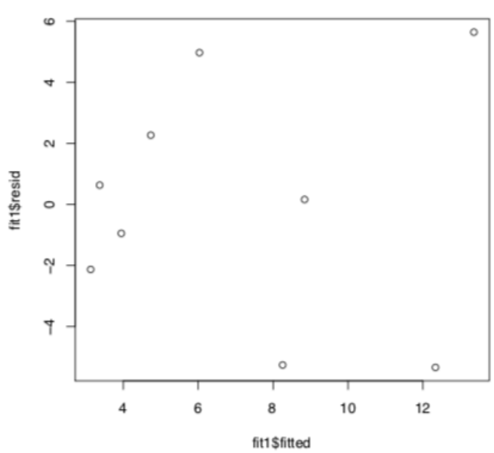


(b) **(1) Simple linear regression without transformation:**

Here is the output. The SSE = 123.5302. The residual plot exhibits increasing variance with fitted values.

```
lm(formula = y ~ x, data = injury)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3351 -2.1281  0.1605  2.2670  5.6382

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1402     3.1412  -0.045   0.9657
x            64.9755    25.1959   2.579   0.0365 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.201 on 7 degrees of freedom
Multiple R-squared:  0.4872,    Adjusted R-squared:  0.4139
F-statistic:  6.65 on 1 and 7 DF,  p-value: 0.03654

> SSE1 = sum((injury$y-fit1$fitted)^2) # SSE
> SSE1
[1] 123.5302
```

The residual plot is as follows.

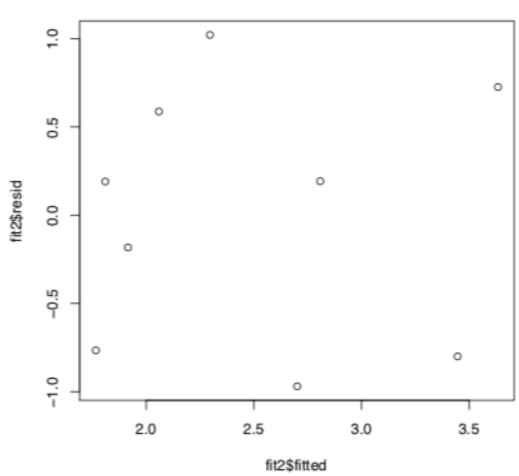


(2) Simple linear regression with square-root transformation:

Here is the output. The SSE = 123.0247, which is slightly less than the previous SSE. The residual plot still exhibits increasing variance with fitted values.

```
Call:
lm(formula = sqrt(y) ~ x, data = injury)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9690 -0.7655  0.1906  0.5874  1.0211 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   1.1692     0.5783   2.022   0.0829 .
x            11.8564     4.6382   2.556   0.0378 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7733 on 7 degrees of freedom
Multiple R-squared:  0.4828,    Adjusted R-squared:  0.4089 
F-statistic: 6.535 on 1 and 7 DF,  p-value: 0.03776

> SSE2 = sum((injury$y-fit2$fitted^2)^2) # SSE
> SSE2
[1] 123.0247
```

The residual plot is as follows.

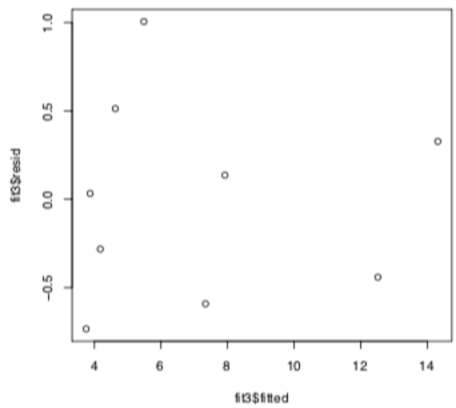


(3) Poisson regression:

Here is the output. The SSE = 117.3472, which is significantly less than the previous SSE. The residual plot does not exhibit increasing variance with fitted values. So this fit is the best of the three fits.

```
Call:
glm(formula = y ~ x, family = poisson(log), data = injury)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.81894  -1.69082   0.06495   1.02407   2.06811

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8945     0.3265   2.739  0.00615 **
x             8.5018     2.1575   3.941 8.13e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 31.859  on 8  degrees of freedom
Residual deviance: 16.291  on 7  degrees of freedom
AIC: 52.251

Number of Fisher Scoring iterations: 5

> SSE3 = sum((injury$y-fit3$fitted)^2) # SSE
> SSE3
[1] 117.3472
```

The residual plot is as follows.



9.4 (Automobile traffic accidents):

(a) The R output of Poisson regression is shown below.

```
Call:
glm(formula = Count ~ Day + Time + Road + Light + Weather
        + Traffic_Control, family = poisson, data = traffic)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-61.305   -3.476   -0.584    1.395   78.294

Coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)                7.444902   0.008396  886.766  < 2e-16
DayWeekend                -0.974418   0.006492 -150.103  < 2e-16
TimeMidday                 0.028733   0.007604    3.779 0.000158
TimeMorning               -0.281834   0.008258  -34.128  < 2e-16
TimeNight                 -0.335986   0.008388  -40.055  < 2e-16
RoadOther                 -2.213312   0.010808 -204.777  < 2e-16
RoadWet                   -1.331213   0.007423 -179.334  < 2e-16
LightDawn/Dusk            -2.048168   0.015995 -128.053  < 2e-16
LightDaylight              0.805835   0.006502  123.930  < 2e-16
LightUnknown              -2.131293   0.016597 -128.416  < 2e-16
WeatherOther              -2.890320   0.014077 -205.327  < 2e-16
WeatherPoor Visibilty     -5.437777   0.049077 -110.802  < 2e-16
WeatherRain/Snow          -1.691636   0.008188 -206.597  < 2e-16
```

```
Traffic_ControlNo Control  0.316084   0.005933   53.275  < 2e-16
Traffic_ControlUnknown    -2.866373   0.019446 -147.404  < 2e-16
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 655308  on 1151  degrees of freedom
Residual deviance: 162264  on 1137  degrees of freedom
AIC: 165864

Number of Fisher Scoring iterations: 7
```

All the coefficients are highly significant but surprisingly many coefficients have counterintuitive signs, e.g., RoadWet and RoadOther have negative signs compared to RoadDry, LightDaylight has a positive sign compared to LightDark, WeatherPoor Visibilty and WeatherRain/Snow have negative signs compared to WeatherClear. Traffic-ControlNo Control has the anticipated sign compared to Traffic-ControlControl since if there are no traffic control signs then one would expect more accidents to occur.

(b) The weighted regression output is shown below. In many cases the coefficients are nearly identical to those obtained with unweighted regression.

```
Call:
glm(formula = Count ~ Day + Time + Road + Light + Weather
    + Traffic_Control,     family = poisson, data = traffic,
    weights = Weight)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-130.483    -6.058    -0.976     2.372   147.782

Coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)                7.398366   0.004095 1806.558  < 2e-16
DayWeekend                -0.974418   0.004196 -232.223  < 2e-16
TimeMidday                 0.014997   0.003697    4.057 4.97e-05
TimeMorning               -0.235902   0.003950  -59.728  < 2e-16
TimeNight                 -0.468529   0.004229 -110.797  < 2e-16
RoadOther                 -2.226426   0.005321 -418.392  < 2e-16
```

```
RoadWet                   -1.321563   0.003619 -365.144  < 2e-16
LightDawn/Dusk            -1.968025   0.007829 -251.392  < 2e-16
LightDaylight              0.909690   0.003246  280.215  < 2e-16
LightUnknown              -2.108363   0.008330 -253.116  < 2e-16
WeatherOther              -2.918760   0.006980 -418.164  < 2e-16
WeatherPoor Visibilty     -5.444327   0.024087 -226.025  < 2e-16
WeatherRain/Snow          -1.683637   0.003992 -421.730  < 2e-16
Traffic-ControlNo Control  0.305629   0.002900  105.392  < 2e-16
Traffic-ControlUnknown    -2.884356   0.009565 -301.567  < 2e-16
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2701803  on 1151  degrees of freedom
Residual deviance:  651854  on 1137  degrees of freedom
AIC: 665102

Number of Fisher Scoring iterations: 7
```