Problem 8.3

$$S = \begin{bmatrix} 4 & 3 \\ 3 & 9 \end{bmatrix}$$

a) $S^{-1} = \frac{1}{36-9}\begin{bmatrix} 9 & -3 \\ -3 & 4 \end{bmatrix} = \begin{bmatrix} 9/27 & -3/27 \\ -3/27 & 4/27 \end{bmatrix} = \begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}$

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

$$L_1 = x_1\left(7/3 - 7/9\right) + x_2\left(-6/9 + 24/27\right) - \begin{bmatrix} \frac{3.5}{2} & -3.5/9 \\ -3/9 & 12/27 \end{bmatrix}\begin{bmatrix} 7 \\ 6 \end{bmatrix}$$

$$\rightarrow L_1 = \frac{14}{9}x_1 + \frac{2}{9}x_2 - \frac{37}{6} \quad \text{for } (5,5) \Rightarrow L_1 = \frac{14}{9}(5) + \frac{2}{9}(5) - \frac{37}{6} = 2.72$$

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

$$L_2 = (4/3 - 4/9)x_1 + (-3/9 + 12/27)x_2 - \begin{bmatrix} 2/3 & -2/9 \\ -1.5/9 & 6/27 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

$$L_2 = \frac{8}{9}x_1 + \frac{1}{9}x_2 - 2 \quad \text{for } (5,5) \Rightarrow L_2 = \frac{8}{9}(5) + \frac{1}{9}(5) - 2 = 3.00$$

$L_1 < L_2 \rightarrow x = (5,5)'$ will be classified to L_2

b) $LD = (7-4, 6-3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (1-1/3)x_1 + (-1/3 + 12/27)x_2$

$$= 2/3\,x_1 + 1/9\,x_2$$

For $x = (5,5)' \rightarrow \frac{2}{3}(5) + \frac{1}{9}(5) = 3.889$

$LD_1 = 2/3(7) + (1/9)(6) = 5.333$, $LD_2 = 7/3(4) + (1/9)(3) = 3$

$\overline{LD} = 4.1665$

since $LD < \overline{LD} \rightarrow$ obs will be classified to Group 2

Problem 8.4

a) Well based on the coefficients, all of the LDF coefficients for CHD are larger than the LDF coefficients for NCHD
→ intuitively (based on historical research) it makes sense that having greater age, blood pressure, and cholesterol lead to increased likelihood of coronary heart disease

b) $L_{CHD} = -28.726 + 0.072(50) + 0.360(95) + 0.079(210)$

$L_{CHD} = 25.664$

$L_{NCHD} = -23.561 + 0.027(50) + .338(95) + .075(210)$

$= 25.649$

c) NCHD → $d_1^2 = (50.0 - 44.81)^2 + (95 - 86.99)^2 + (210 - 201.27)^2$

$d_1 = 12.935$

CHD → $d_2^2 = (50 - 56.86)^2 + (95 - 95.62)^2 + (210 - 221.51)^2$

→ $d_2 = 13.414$

→ Since $d_{CHD} = 13.414 > d_{NCHD} = 12.935$, classify observation as NCHD.

→ The reason this differs from part (b) is because ~~the~~ euclidean distance does not the correlations ~~among~~ among variables and weighs all variables equally. The LDF calculation uses the covariance matrix

b) $Age^* = 50 \quad DBP = 95 \quad CHL = 210$

$L_{CHD}(x^*) = .072 \times 50 + .36 \times 95 + .079 \times 210 - 28.726 = 25.664$

$L_{NCHD}(x^*) = .027 \times 50 + .338 \times 95 + .075 \times 210 - 23.561 = 25.649$

Substract 25 from both $\Rightarrow L_{CHD} = .664$ & $L_{NCHD} = .649$

$$\hat{P}^*_{CHD} = \frac{exp(.664)}{exp(.664) + exp(.649)} = \frac{1.944}{1.943 + 1.944} = \frac{1.944}{3.887} = 49.62\%$$

$$\hat{P}^*_{NCHD} = \frac{exp(.649)}{exp(.664) + exp(.649)} = \frac{1.943}{3.887} = 50.37\%$$

Since $\hat{P}^*_{NCHD} > \hat{P}^*_{CHD} \Rightarrow$ Classify as NCHD

MSiA_401_HW8

Parth Patel, Jieda Li, Kris Nikolov, Kristiyan Dimitrov

12/2/2019

MSiA 401 Homework 8

Problem 8.5

```
library(MASS)
iris_df = read.csv("Iris.csv")
fit=lda(Species_No ~ Petal_width+Petal_length+Sepal_width+Sepal_length, data=iris_df)

#part A answer:
fit
```

```
## Call:
## lda(Species_No ~ Petal_width + Petal_length + Sepal_width + Sepal_length,
##     data = iris_df)
##
## Prior probabilities of groups:
##         1         2         3
## 0.3333333 0.3333333 0.3333333
##
## Group means:
##   Petal_width Petal_length Sepal_width Sepal_length
## 1       0.246        1.462       3.428        5.006
## 2       1.326        4.260       2.770        5.936
## 3       2.026        5.552       2.974        6.588
##
## Coefficients of linear discriminants:
##                      LD1         LD2
## Petal_width  -2.8104603 -2.83918785
## Petal_length -2.2012117  0.93192121
## Sepal_width   1.5344731 -2.16452123
## Sepal_length  0.8293776 -0.02410215
##
## Proportion of trace:
##    LD1    LD2
## 0.9912 0.0088
```

```
#part b answer:
predict(fit,newdata=data.frame(Sepal_length = 5.5, Sepal_width = 3.0, Petal_length = 4.0, Petal_
width = 1.5))$posterior
```

```
##              1         2           3
## 1 1.962381e-19 0.9993753 0.0006246577
```

```
#We see that Versicolor has the highest probability therefore we would classify this observation as Versicolor
```

Problem 9.3

```
df_airinjur = read.csv("Airline-Injury.csv")

fit_lr = lm(y~x,data=df_airinjur)
lr_anova = anova(fit_lr)
print(paste0("SSE for linear regression: ", lr_anova$`Sum Sq`[2]))
```

```
## [1] "SSE for linear regression: 123.530182671431"
```

```
fit_lr_sqt = lm(sqrt(y)~x,data=df_airinjur)
sse2 = sum((df_airinjur$y - (fit_lr_sqt$fitted.values)^2)^2)
print(paste0("SSE for linear regression w/sqrt transform of y: ", sse2))
```

```
## [1] "SSE for linear regression w/sqrt transform of y: 123.024663801919"
```

```
fit_poi=glm(y~x, data=df_airinjur, family=poisson(log))
fitted_Vals = fit_poi$fitted.values
residuals = df_airinjur$y-fitted_Vals
print(paste0("SSE for poisson regression: ",sum(residuals^2)))
```

```
## [1] "SSE for poisson regression: 117.34718250811"
```

We see that the SSE for the sqrt() transformed y appears to be only slightly lower than the regular lm model. Furthermore, the variables are still reasonably significant. The plot below clearly shows that a linear fit is not the best idea. This is also supported by the low R^2 we get for both models.

I would go with poisson regression applied to the sqrt of y transformation as its SSE is the lowest among the three. Also the variables are more significant and the overall model deviance passes the chi-squared significance test

Problem 9.4

```
df_crashdale = read.csv("crashdata2014-summary.csv")
fit_unwgt_poi = glm(formula = Count ~ Day + Time + Road + Light + Weather + TrafficControl, fami
ly = poisson, data = df_crashdale)
summary(fit_unwgt_poi)
```

```
##
## Call:
## glm(formula = Count ~ Day + Time + Road + Light + Weather + TrafficControl,
##     family = poisson, data = df_crashdale)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -48.765   -4.197   -1.071    1.111   65.305
##
## Coefficients:
##                           Estimate Std. Error  z value Pr(>|z|)
## (Intercept)               7.103091   0.010011  709.518   <2e-16 ***
## DayWeekend               -0.966953   0.007781 -124.266   <2e-16 ***
## TimeMidday               -0.008045   0.009227   -0.872    0.383
## TimeMorning              -0.222000   0.009763  -22.738   <2e-16 ***
## TimeNight                -0.335498   0.010084  -33.270   <2e-16 ***
## RoadOther                -2.273723   0.013582 -167.410   <2e-16 ***
## RoadWet                  -1.134737   0.008412 -134.898   <2e-16 ***
## LightDawn/Dusk           -2.034986   0.019172 -106.142   <2e-16 ***
## LightDaylight             0.796045   0.007852  101.385   <2e-16 ***
## LightUnknown             -1.784394   0.017190 -103.806   <2e-16 ***
## WeatherOther             -2.587845   0.014941 -173.200   <2e-16 ***
## WeatherPoor Visibilty    -3.598460   0.024209 -148.644   <2e-16 ***
## WeatherRain/Snow         -1.665179   0.009906 -168.093   <2e-16 ***
## TrafficControlNo Control  0.089846   0.007090   12.671   <2e-16 ***
## TrafficControlUnknown    -2.533891   0.018901 -134.063   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 428848  on 1151  degrees of freedom
## Residual deviance: 123720  on 1137  degrees of freedom
## AIC: 127219
##
## Number of Fisher Scoring iterations: 7
```

```
fitted_Vals = fit_unwgt_poi$fitted.values
residuals = df_crashdale$Count-fitted_Vals
print(paste0("SSE for unweighted poisson regression: ",sum(residuals^2)))
```

```
## [1] "SSE for unweighted poisson regression: 65068375.2567379"
```

Comments on coefficients:

- DayWeekend has negative coefficient, which makes sense - fewer people drive on weekend –> fewer crashes
- TimeMidday is not signicicant –> no real difference b/w Midday & Evening, which makes sense , because the counts are almost the same (23,399 & 23,588)
- Time Morning & Time Night - negative coefficients make sense, because there are fewer people driving in morning & night than in evening (which is the reference category)

- NoTrafficControl - positive coefficient makes sense - peple are more likely to speed and therefore crash with no traffic control
- RoadWet - negative coefficient doesn't make much sense - would expect poor road conditions to increase crashes
- LightDawn/Dusk - negative coefficient compared to reference category Dark doesn't make much sense. This variable is very correlated with Time, which probably explains the coefficients i.e. signs are the same as Time Morning/Night
- LightDaylight - same as above
- WeatherPoor Visibility & Rain Snow - correlated with RoadWet and the Light variables.

```
fit_wgt_poi = glm(formula = Count ~ Day + Time + Road + Light + Weather + TrafficControl, family
= poisson, data = df_crashdale, weights = Weight)
summary(fit_wgt_poi)
```

```
##
## Call:
## glm(formula = Count ~ Day + Time + Road + Light + Weather + TrafficControl,
##     family = poisson, data = df_crashdale, weights = Weight)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -103.052    -6.938    -1.792     2.190   133.894
##
## Coefficients:
##                            Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                7.061573   0.004882 1446.573  < 2e-16 ***
## DayWeekend                -0.966953   0.005027 -192.353  < 2e-16 ***
## TimeMidday                -0.030746   0.004488   -6.851 7.31e-12 ***
## TimeMorning               -0.179012   0.004666  -38.365  < 2e-16 ***
## TimeNight                 -0.481346   0.005095  -94.477  < 2e-16 ***
## RoadOther                 -2.281779   0.006684 -341.363  < 2e-16 ***
## RoadWet                   -1.116644   0.004097 -272.569  < 2e-16 ***
## LightDawn/Dusk            -1.961127   0.009426 -208.060  < 2e-16 ***
## LightDaylight              0.904702   0.003923  230.590  < 2e-16 ***
## LightUnknown              -1.757028   0.008629 -203.623  < 2e-16 ***
## WeatherOther              -2.610596   0.007394 -353.064  < 2e-16 ***
## WeatherPoor Visibilty     -3.568698   0.011684 -305.443  < 2e-16 ***
## WeatherRain/Snow          -1.655916   0.004832 -342.703  < 2e-16 ***
## TrafficControlNo Control   0.067936   0.003468   19.590  < 2e-16 ***
## TrafficControlUnknown     -2.567926   0.009342 -274.886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1764135  on 1151  degrees of freedom
## Residual deviance:  496975  on 1137  degrees of freedom
## AIC: 509909
##
## Number of Fisher Scoring iterations: 7
```

```
SSE_wgtd = sum((df_crashdale$Count-(fit_wgt_poi$fitted)*(df_crashdale$Weight))^2)
print(paste0("SSE for weighted poisson regression: ", SSE_wgtd))
```

```
## [1] "SSE for weighted poisson regression: 1096273153.70111"
```

In the weighted regression model all of the predictors are significant whereas there was a predictor that was not significant in the unweighted approach But the AIC and deviance are much larger for the weighted approach vs the unweighted approach. The SSE got much larger for the weighted approach as well.