HW 8 | Shawn Li

Problem 1

Discriminant analysis:

```
Call:
lda(CC ~ SSPG + IR + RW, data = diabetes)

Prior probabilities of groups:
        1         2         3
0.2275862 0.2482759 0.5241379

Group means:
      SSPG       IR        RW
1 318.8788 106.0000 0.9839394
2 208.9722 288.0000 1.0558333
3 114.0000 172.6447 0.9372368

Coefficients of linear discriminants:
              LD1          LD2
SSPG -0.016193147 -0.002202638
IR    0.004058614 -0.007450127
RW    2.393055796 -3.831511854

Proportion of trace:
   LD1    LD2
0.8238 0.1762
```

```
> classification.table.prob1;
   predict.lda.prob1
     1  2  3
1 26  4  3
2  0 22 14
3  2  4 70
```

The number of miss classifications is 27. The misclassification rate is 0.1862.

Multinomial logistic regression: (Result from HW7 Prob4)

```
Call:
mlogit(formula = CC ~ 0 | IR + SSPG + RW, data = diab, reflevel = "3",
    method = "nr", print.level = 0)

Frequencies of alternatives:
      3       1       2
0.52414 0.22759 0.24828

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 0.00028
successive function values within tolerance limits

Coefficients :
               Estimate Std. Error t-value  Pr(>|t|)
1:(intercept) -1.8446132  3.4634601 -0.5326  0.594316
2:(intercept) -7.6154166  2.3356317 -3.2605  0.001112 **
1:IR          -0.0133537  0.0050193 -2.6605  0.007804 **
2:IR           0.0035868  0.0023492  1.5268  0.126803
1:SSPG         0.0455039  0.0092415  4.9239 8.486e-07 ***
2:SSPG         0.0164141  0.0049819  3.2948  0.000985 ***
1:RW          -5.8674627  3.8665785 -1.5175  0.129145
2:RW           3.4727694  2.4461624  1.4197  0.155701
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -68.415
McFadden R^2:  0.53805
Likelihood ratio test : chisq = 159.37 (p.value = < 2.22e-16)
```

```
> ctable;
     Y.hat
        1   2   3 Sum
  1    27   3   3  33
  2     0  24  12  36
  3     2   5  69  76
  Sum  29  32  84 145
> sum(diag(ctable)[-4])/diag(ctable)[4]
      Sum
0.8275862
```

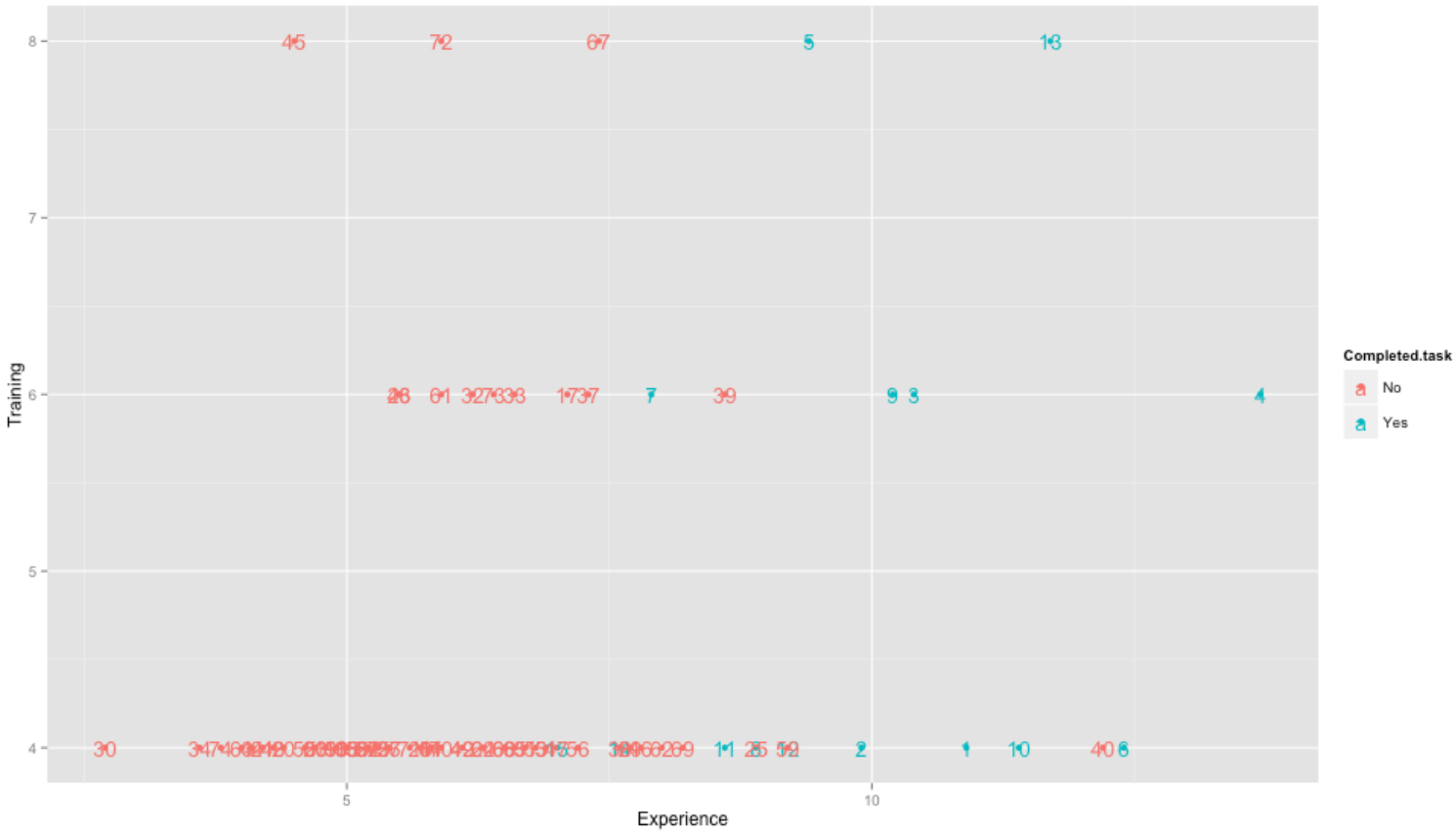
The number of miss classifications is 25. The misclassification rate is 0.1724.

Result of multinomial logistic regression is **slightly better** than result of discriminant analysis since the misclassification rate is lower for multinomial logistic regression.

HW 8 | Shawn Li

Problem2

(a)



Based on the labeled scatter plot, they can be well-separated by a straight line.

(b)

```
Call:
lda(Completed.task ~ Training + Experience, data = systemAdmin)

Prior probabilities of groups:
 No Yes
0.8 0.2

Group means:
    Training Experience
No  4.500000   6.013333
Yes 5.066667   9.946667

Coefficients of linear discriminants:
                 LD1
Training   0.06009381
Experience 0.60173656
```

Result of discriminant analysis is above. Discriminant function scores and predicted probabilities to classify the75 observations are shown at next page.

HW 8 | Shawn Li

	discriminant function scores LD1		discriminant function scores LD1
1	2.43026237	41	-0.63859410
2	1.82852581	42	-0.45807313
3	2.24958171	43	-0.69892745
4	4.23531237	44	-1.66154626
5	1.76803277	45	-1.18047639
6	3.33286722	46	0.56487903
7	0.74524030	47	0.08348978
8	1.22678925	48	-1.54119894
9	2.12923440	49	-0.45807313
10	2.73113065	50	-1.18015701
11	1.04626828	51	0.02331612
12	1.40731022	52	-0.99963604
13	3.15202686	53	-0.03685754
14	0.44453171	54	-0.63859410
15	0.08348978	55	-1.36067798
16	-1.18015701	56	0.20383709
17	0.26385105	57	-0.99963604
18	-1.11998335	58	-1.05980969
19	-1.24033066	59	1.40731022
20	-1.48102529	60	-1.30050432
21	-1.60137260	61	-0.45823282
22	-0.33772582	62	0.68522634
23	-1.30050432	63	-1.72171991
24	-0.99963604	64	-0.33772582
25	1.22678925	65	-0.15720485
26	-0.69892745	66	-0.21737851
27	-0.87928873	67	0.56455964
28	-0.69876776	68	-0.21737851
29	0.50470537	69	0.80557365
30	-2.50397745	70	-0.57842044
31	-0.09703119	71	-0.75894141
32	-0.27771186	72	-0.33804521
33	-0.03701723	73	-0.15736454
34	-1.96241454	74	-1.84206723
35	-1.11998335	75	-0.93946238
36	-0.87928873		
37	0.38419836		
38	0.44453171		

	predicted probabilities No	Yes		predicted probabilities No	Yes
1	0.061849281	0.9381507186	41	0.990517290	0.0094827101
2	0.218486079	0.7815139215	42	0.985447705	0.0145522953
3	0.092337951	0.9076620489	43	0.991785537	0.0082144629
4	0.000864128	0.9991358720	44	0.999179476	0.0008205242
5	0.244295124	0.7557048762	45	0.997400293	0.0025997068
6	0.007493041	0.9925069593	46	0.853129264	0.1468707360
7	0.790233222	0.2097667778	47	0.948588724	0.0514112760
8	0.542444802	0.4575551975	48	0.998904892	0.0010951082
9	0.119573312	0.8804266885	49	0.985447705	0.0145522953
10	0.031021561	0.9689784387	50	0.997398304	0.0026016959
11	0.646479188	0.3535208122	51	0.955194657	0.0448053432
12	0.434572072	0.5654279281	52	0.995992519	0.0040074811
13	0.011520012	0.9884799879	53	0.960986690	0.0390133098
14	0.885775540	0.1142244600	54	0.990517290	0.0094827101
15	0.948588724	0.0514112760	55	0.998311789	0.0016882105
16	0.997398304	0.0026016959	56	0.932527460	0.0674725395
17	0.922876783	0.0771232171	57	0.995992519	0.0040074811
18	0.996995145	0.0030048545	58	0.996529731	0.0034702692
19	0.997747494	0.0022525065	59	0.434572072	0.5654279281
20	0.998734898	0.0012651023	60	0.998049908	0.0019500923
21	0.999052065	0.0009479349	61	0.985453202	0.0145467981
22	0.980666772	0.0193332283	62	0.813120962	0.1868790382
23	0.998049908	0.0019500923	63	0.999289774	0.0007102265
24	0.995992519	0.0040074811	64	0.980666772	0.0193332283
25	0.542444802	0.4575551975	65	0.970487871	0.0295121286
26	0.991785537	0.0082144629	66	0.974356011	0.0256439895
27	0.994657140	0.0053428603	67	0.853225319	0.1467746807
28	0.991782413	0.0082175871	68	0.974356011	0.0256439895
29	0.870324186	0.1296758142	69	0.765213526	0.2347864737
30	0.999891356	0.0001086440	70	0.989059546	0.0109404543
31	0.966056588	0.0339434115	71	0.992879964	0.0071200355
32	0.977737137	0.0222628633	72	0.980681305	0.0193186953
33	0.961001062	0.0389989379	73	0.970498851	0.0295011494
34	0.999601377	0.0003986229	74	0.999467905	0.0005320951
35	0.996995145	0.0030048545	75	0.995372531	0.0046274695
36	0.994657140	0.0053428603			
37	0.899632823	0.1003671773			
38	0.885775540	0.1142244600			
39	0.578115515	0.4218844846			

HW 8 | Shawn Li

```
> classification.table.prob2;
     predict.lda.prob2
      No Yes
  No  58   2
  Yes  5  10
 > classification.rate.prob2 = sum(diag(classification.table.prob2)[1:2])/n;
 > classification.rate.prob2
 [1] 0.9066667
 > misclassification.rate.overall = 1-classification.rate.prob2;
 > misclassification.rate.overall
 [1] 0.09333333
 > mislcassification.rate.YES = 5/n
 > mislcassification.rate.YES
 [1] 0.06666667
 > mislcassification.rate.NO = 2/n
 > mislcassification.rate.NO
 [1] 0.02666667
```

Following table gives misclassification rate for over and each group.

	Overall	Group YES	Group No
Misclassification rate	0.09333333	0.333	0.0333

(c)

```
$class
[1] No
Levels: No Yes

$posterior
         No          Yes
1 0.9990524 0.0009475719

$x
        LD1
1 -1.601532
```

The systems administrator **will not** complete tasks. It is classified to group Completed_tasks = No

(d)

Prior probabilities of group: No: 60/75 ; Yes: 0.2. Under the prior probabilities, we got a classification table, which is in (b):

```
     predict.lda.prob2.d
      No Yes
  No  58   2
  Yes  5  10
```

The posterior probability is conditional probability density function for x being a member of group j.
Posterior probability of group: No: 63/75; Yes: 12/75.

HW 8 | Shawn Li

```
Call:
lda(Completed.task ~ Training + Experience, data = systemAdmin,
    prior = c(63/75, 12/75))

Prior probabilities of groups:
  No  Yes
0.84 0.16

Group means:
    Training Experience
No  4.500000   6.013333
Yes 5.066667   9.946667

Coefficients of linear discriminants:
                 LD1
Training   0.06009381
Experience 0.60173656
```

```
predict.lda.prob2.d
     No Yes
No   59   1
Yes   6   9
```

I got a new classification table. The overall misclassification rate is 0.0933, which is same to (b). So, **overall misclassification rate does not change**.

	Overall	Group YES	Group No
Misclassification rate	0.09333333	0.4	0.0166

Comparing misclassification rate for each group between (b) and (d). For group Yes, the misclassification rate goes up from 0.333 to 0.4; for group No, the misclassification rate goes down from 0.033 to 0.0166.

Problem 3

```
> summary(fit.poisson)

Call:
glm(formula = Y ~ N, family = "poisson", 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 **
N             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
```

```
> summary(fit.lq)

Call:
lm(formula = Y ~ N, 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
N            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
```

HW 8 | Shawn Li

```
> summary(fit.tlq)

Call:
lm(formula = sqrt(Y) ~ N, 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 .
N            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
```

```
> SSE.poisson
[1] 117.3472
> SSE.lq
[1] 123.5302
> SSE.tlq
[1] 393.9632
```

Summary of Poisson regression, Least Squares fit (lq) and Transformed Least Squares fit (tlq) are shown above. For poisson fit, predictor N is highly significant; for lq fit and tlq fot, N is still significant in these two models. To compare these three models, I get their SSE: SSE for poisson is 117.3472 and SSE for Least Squares fit is 123.5302. Since response variable Y is a count variable, which poisson regression is good at. SSE(poisson) < SSE(lq) tells Poission models fits the data better than does the linear regression model.

After transformed Y into sqrt(Y), the response variable has a normal distribution. SSE(tlq) = 393.9632 is far larger than SSE(Poisson)=117.3472. The SSE indicates that the poisson model still fits the data better.

Therefore, Poisson model provides the best description of the date since it has the smallest SSE.

Problem 4

$$f(y;\mu) = \frac{1}{\mu}\exp\left(-\frac{y}{\mu}\right) = \exp\left(\ln\left(\frac{1}{\mu}\right)\right)\exp\left(-y*\frac{1}{\mu}\right) = \exp\left(-\ln(\mu) - y*\frac{1}{\mu}\right)$$

So I can get: $a(y) = -y, b(\mu) = \frac{1}{\mu}, c(y) = 0, d(\mu) = -\ln(\mu)$. It belongs to the exponential family.

Natural link function for this distribution: $g(\mu) = b(\mu) = \frac{1}{\mu}$.