Answer Key for Problem Set #3

(1) Get familiar with the dataset.

(a)

```
    Variable |      Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     depress |     2938    .0646698    .2459842          0          1
    appetite |     2938    .1109598    .3141359          0          1
       sleep |     2938    .1364874    .3433638          0          1
     fatigue |     2938    .0918993    .2889329          0          1
    concentr |     2938    .0636487    .2441677          0          1
      morbid |     2938    .1211709    .3263813          0          1
```

From this output, we can tell the prevalence of each symptom. The most prevalent symptom is sleeping problems with almost 14% of the sample reporting insomnia or hypersomnia. Suicidal or morbid thoughts is also relatively prevalent with about 12% of the sample reporting these symptoms. The least prevalent symptoms are concentration problems and depressed mood, with slightly over 6% of the sample reporting these symptoms.

(b) Table of cross tabs (y/y is in upper left hand corner, etc)

	Depress		Appetite		Sleep		Fatigue		Thoughts		Suicide
D											
+A	85	241									
	105	2507									
Sl	99	302	134	267							
	91	2446	192	2345							
F	84	186	95	175	127	143					
	106	2562	132	2437	274	2394					
T	70	117	75	112	101	86	89	98			
	120	2631	251	2500	300	2451	181	2570			
Su	97	259	106	250	147	209	95	261	87	269	
	93	2489	220	2362	254	2328	175	2407	100	2482	

(c) There are $2^6 = 64$ possible response patterns. I do not expect to see them all and this is because all of the symptoms are relatively rare. Hence, I would expect that some of the patterns with more than 2 symptoms would not be reported in this sample. The majority of individuals (68%) report no symptoms and very few report more than 3 symptoms (<4%). We can see the breakdown of number of reported symptoms in the table below:

```
. egen tot=rsum(dep app sle fat conc morb)
. tab tot
        tot |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |       2029       69.06       69.06
          1 |        487       16.58       85.64
          2 |        205        6.98       92.61
          3 |        105        3.57       96.19
          4 |         61        2.08       98.26
          5 |         32        1.09       99.35
```

```
          6 |          19        0.65      100.00
------------+-----------------------------------
      Total |        2938      100.00
```

All of the observed patterns may be seen in the following table (generated is SAS 8.12):

Obs	appetite	concentr	depress	fatigue	morbid	sleep	COUNT	PERCENT
1	0	0	0	0	0	0	2029	69.0606
2	0	0	0	0	1	0	127	4.3227
3	0	0	0	0	0	1	122	4.1525
4	1	0	0	0	0	0	113	3.8462
5	0	0	0	1	0	0	68	2.3145
6	0	0	0	0	1	1	35	1.1913
7	1	0	0	0	0	1	31	1.0551
8	0	1	0	0	0	0	30	1.0211
9	0	0	0	1	0	1	27	0.9190
10	0	0	1	0	0	0	27	0.9190
11	1	1	1	1	1	1	19	0.6467
12	1	0	0	0	1	0	18	0.6127
13	0	0	1	0	1	0	14	0.4765
14	1	0	0	0	1	1	13	0.4425
15	1	0	0	1	0	0	13	0.4425
16	0	1	0	0	0	1	11	0.3744
17	0	1	0	1	0	1	10	0.3404
18	0	1	1	1	1	1	10	0.3404
19	1	0	0	1	0	1	10	0.3404
20	0	0	0	1	1	0	9	0.3063
21	0	0	1	0	0	1	9	0.3063
22	1	0	1	0	1	1	9	0.3063
23	0	0	1	0	1	1	8	0.2723
24	0	1	0	0	1	1	8	0.2723
25	1	0	1	0	0	0	8	0.2723
26	0	1	0	1	0	0	7	0.2383
27	0	1	0	1	1	1	7	0.2383
28	1	0	1	1	0	1	7	0.2383
29	1	1	0	0	0	0	7	0.2383
30	1	1	1	0	1	1	7	0.2383
31	0	0	0	1	1	1	6	0.2042
32	0	0	1	1	0	0	6	0.2042
33	0	0	1	1	1	1	6	0.2042
34	0	1	0	0	1	0	6	0.2042
35	1	0	0	1	1	0	6	0.2042
36	1	1	0	0	1	1	6	0.2042
37	0	1	0	1	1	0	5	0.1702
38	0	1	1	0	1	0	5	0.1702
39	0	1	1	1	0	0	5	0.1702
40	1	0	1	0	0	1	5	0.1702
41	1	0	1	1	1	0	5	0.1702
42	1	1	0	0	0	1	5	0.1702
43	0	0	1	1	0	1	4	0.1361
44	0	1	1	0	0	0	4	0.1361
45	1	0	0	1	1	1	4	0.1361
46	1	0	1	1	1	1	4	0.1361
47	1	1	0	1	1	0	4	0.1361
48	1	1	0	1	1	1	4	0.1361
49	1	1	1	1	0	1	4	0.1361
50	0	0	1	1	1	0	3	0.1021
51	1	0	1	1	0	0	3	0.1021
52	1	1	0	1	0	0	3	0.10211
53	1	1	0	1	0	1	3	0.10211
54	1	1	1	0	0	1	3	0.10211
55	1	1	1	1	0	0	3	0.10211
56	1	1	1	1	1	0	3	0.10211
57	0	1	1	1	0	1	2	0.06807
58	1	0	1	0	1	0	2	0.06807
59	1	1	1	0	0	0	2	0.06807
60	0	1	1	0	0	1	1	0.03404
61	0	1	1	0	1	1	1	0.03404
62	1	1	0	0	1	0	1	0.03404
63	1	1	1	0	1	0	1	0.03404

(2) Fit LC models with 1, 2 and 3 class models.

(See attached)

(3)

Variable	1-Class Model	2-Class Model		3-Class Model		
	Class 1	Class 1	Class 2	Class 1	Class 2	Class 3
Depress	0.06	0.01	0.38	0.01	0.21	0.81
Appetite	0.11	0.05	0.46	0.04	0.32	0.71
Sleep	0.14	0.06	0.62	0.04	0.47	0.83
Fatigue	0.09	0.03	0.46	0.02	0.30	0.76
Concentration	0.06	0.01	0.37	0.01	0.21	0.75
Morbid	0.12	0.06	0.50	0.05	0.34	0.80
Class Size	1.00	0.86	0.14	0.79	0.18	0.03

(4) Consider the precision estimates of the parameters. Is there evidence that any of the models are not identifiable/estimable?

The two and three class models appear identifiable. The standard errors are all relatively small. To be identifiable, the condition $M-1+(M*K) < 2^K - 1$ must be satified, where M is the number of classes, and K is the number of classes (though this does not guarantee it).

2-class model : $13 < 63$; this model may be identifiable
3-class model : $20 < 63$; this model may be identifiable

(5) Interpret the estimable models.

In the two class model, we have a class of "normals," which comprises 86% of the population and a class of "depressed" individuals, which comprises the remaining 14% of the population. The normals report symptoms relatively rarely. Specifically, an individual from the normal class has less than a 6% chance of reporting each of the symptoms. The depressed class is more likely to report each of the symptoms. For example, an individual from the depressed class has at least a 37% chance of reporting each of the symptoms. The symptom prevalences in this class range from 0.37 to 0.62.

The three class model has a class of normals (79%), a class of severely depressed individuals (3%), and a "subclinically" depressed class (18%). The normal class is very similar to that in the 2 class model–all symptoms are reported relatively rarely (5% in this case for each of the symptoms). The severely depressed class reports symptoms quite frequently: an individual in the severely depressed class has at least a 75% chance of reporting each of the symptoms, with the symptom prevalences ranging from 0.75 to 0.83. The "subdromal" depression class is in between the normals and the severely depressed in the sense that the symptom prevalences are all larger than those for the normals, but smaller than those for the severely depressed. The range of symptom prevalences in this class is from 0.21 to 0.47.

(6) Calculate the posterior probabilities of class membership for each class for the 2 and 3 class models. (I go through the math for reporting no symptoms in the two class model on the next page and provide just the answers for the others).

2 class model:
(a) reporting no symptoms
P(class=1|000000)=0.996
P(class=2|000000)=0.004
(b) reporting all symptoms
P(class=1|111111)<0.001
P(class=2|111111)>0.999

3 class model:
(a) reporting no symptoms
P(class=1|000000)=0.974
P(class=2|000000)=0.026
P(class=3|000000)<0.001
(b) reporting all symptoms
P(class=1|111111)<0.001
P(class=2|111111)=0.019
P(class=3|111111)=0.977

For 2 class model: posterior probabilities of class membership for an individual reporting no symptoms (i.e. P(class=1|000000) and P(class=2|000000))

$$P(class=1\,|\,000000)=\frac{P(000000\,|\,class=1)P(class=1)}{P(000000\,|\,class=1)P(class=1)+P(000000\,|\,class=2)P(class=2)}$$

$$=\frac{0.798\times 0.86}{0.798\times 0.86+0.021\times 0.14}=0.996$$

$$P(class=2\,|\,000000)=\frac{P(000000\,|\,class=2)P(class=2)}{P(000000\,|\,class=2)P(class=2)+P(000000\,|\,class=1)P(class=1)}$$

$$=\frac{0.021\times 0.14}{0.021\times 0.14+0.798\times 0.86}=0.004$$

where, by conditional independence,

$$P(000000\,|\,class=1)=P(dep=0\,|\,class=1)P(app=0\,|\,class=1)P(slp=0\,|\,class=1)P(\,fat=0\,|\,class=1)P(co$$

$$=0.99\times 0.95\times 0.94\times 0.97\times 0.99\times 0.94=0.798$$

$$P(000000\,|\,class=2)=P(dep=0\,|\,class=2)P(app=0\,|\,class=2)P(slp=0\,|\,class=2)P(\,fat=0\,|\,class=2)P(c$$

$$=0.61\times 0.54\times 0.38\times 0.54\times 0.63\times 0.50=0.021$$

(7) How many individuals are predicted to report no symptoms by the 3 class model?

Using the information from the previous question, we have estimated that
P(000000)=P(000000|class=1)P(class=1) + P(000000|class=2)P(class=2) + P(000000|class=3)P(class=3)
=

(0.841)(0.79) + (0.102)(0.18) + (0.00011)(0.03) = 0.683

This means that we estimate that 68.3% of the sample would give no symptoms as his/her response pattern. So, in a sample of size 2938, we estimate that (2938)(0.683) = 2006.6 individuals would report this pattern. As is noted, 2029 did report this pattern, which is fairly close to 2006.6 indicating that the model describes the prevalence of this pattern quite well. This provides some evidence about goodness of fit, but we would need more information about how close the observed and predicted frequencies are for other patterns.

How many individuals are predicted to report all symptoms by the 3 class model?

Using the same calculation as above, we predict that 0.66% of the sample report all symptoms. This

corresponds to 19.2 individuals which is consistent with the 19 individuals that we observed to report this pattern. Hence, both of the predicted frequencies of response the patterns in this question are consistent with the observed data, suggesting good fit for the 3 class model. As noted above, however, we would want to consider more than just two of the 64 possible patterns.

How about in the 2 class model?

In the two class model, 68.9% of the sample is predicted to have no symptoms, which corresponds to 2025.1 individuals. Only 0.13% of the sample is predicted to report all symptoms, which is 3.9 individuals. We can see in comparing these to the observed frequencies for these patterns (2029 and 19, respectively) that while the three class model did a better job of predicting those reporting no symptoms, the two class model did a better job of predicting those reporting all symptoms.

(8) Which model is most appropriate?

Looking at the fit statistics, the AIC, the BIC, and the Chi-squared both favor the three class model. We should also rely on the clinical evidence that we have. One thing we may consider is to look at the DSM criteria and see if these results describe a class that is similar to that in the DSM criteria. It turns out that the "severe class" in the three class model looks similar to the DSM diagnosis of depression. Using that additional information, I would then choose the three class model as most appropriate.

(9) The Chi-squared statistic is not valid in this case. We can see from the Sas table above, which shows all of the patterns observed, that some of the patterns have fewer than 5 counts in the cells. And so, we can conclude that the 64^{th} pattern would have no observations, which would violate the assumption of "large cell counts" needed to use the Chi-square as a goodness of fit statistic.

(10) Does the LC model help us understand depression?

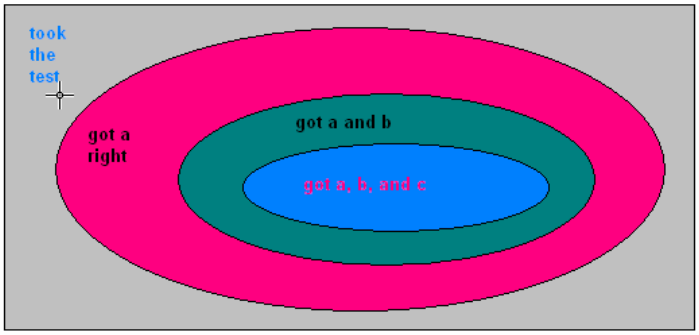
Yes, it does. It simplifies our understanding of depression from 64 symptom patterns to 3 (or 2) classes of depression.
I think that this is an appropriate application of the LC model. We are trying to "diagnose" individuals into classes of depression. In other words, we are trying to "cluster" individuals and not symptoms as one would do in a factor analysis. The classes that we have found in the 3 class model are consistent with "expert opinion" in the sense that one of the classes agrees quite well with the DSM criteria.

(11) Guttman scaling would not be a good approach in this case. There does not appear to be a hierarchy of symptoms as can be seen in the 3 class model and can be seen in the crosstabs that we performed in question 1.

One example where this could be helpful is in the development of computer adaptive tests, like the GRE. Questions could be as follows:

a) 1 + 1
b) 2 * 4
c) ½ + 1/3
d) cos 45
e) $f(x) = x^2$; $f'(x) =$

and so on… Presumably the relationships between groups of people who answered the questions correctly would look like this:



The items would be ordered in the order of the most inclusive to the least inclusive

Extra Credit:

a) The restriction “if a person is suicidal, they must be in the most severe class” means that all of the suicidal people are restricted to the most severe class, but non-suicidal people may also be in the most severe class. However, the restriction “if they are in the most severe class, they must be suicidal” means that suicidal people may be in any class, but only suicidal people may be in the most severe class.

b) “If a person is suicidal, they must be in the most severe class” : P(most severe class | suicidal) = 1.
“If they are in the most severe class, they must be suicidal”: P(suicidal | most severe class) = 1.

c) “If they are in the most severe class, they must be suicidal”: for %C#1% use [morbid$1@-100]. That is, constrain the conditional probability of being suicidal in class 1 to 1.
“If a person is suicidal, they must be in the most severe class”: for %C#2% and %C#3% use [morbid$1@100]. That is, constrain the conditional probabilities of being suicidal in classes 2 and 3 to 0.

d) I would constrain the model if I had a strong scientific justification to do so, for example if I were trying to confirm a diagnostic criteria. I would not constrain the model if I was unsure of the justification, or if I was unfamiliar with the study population.

e) I would check the source document to determine whether or not a data entry/programming error occurred. I would also talk to subject-matter experts to see if there is a good scientific reason for the response patterns I intend to constrain.