STAT505 Assessment #10

1. A data set includes six variables that measure the wellbeing of patients undergoing radiotherapy. The variables are *symptoms* (number of side effect symptoms), activity (amount of daily activity on a 1-5 scale), sleep (quality of sleep on 1-5 scale), eat (food consumed on 1-3 scale), appetite (on a 1-5 scale), and skinreact (measure of skin reaction on 0-3 scale). Following are factor loadings from a factor analysis of the standardized variables. The principal components method was used along with a varimax rotation.

	Factor1	Factor2	Factor3
symptoms	0.763	0.239	0.133
activity	0.899	0.036	-0.057
sleep	0.061	0.898	0.146
eat	0.562	0.556	-0.034
appetite	0.618	0.650	-0.109
skinreact	0.016	0.067	0.985

   (a) Calculate the communality for the symptoms variable. Write a sentence that interprets this value.

   This communality is defined as $.763^2 + .239^2 + .133^2 = .657$. This is interpreted as the percent of the variability in symptoms explained by these three factors.

   (b) Calculate the specific variance for the symptoms variable. *Recall we're using standardized variables.*

   The total variance for symptoms is 1 (since its standardized), which is the sum of its communality and specific variance. By subtraction, the specific variance is $1 - .657 = .343$.

   (c) The total variance explained by a factor is the sum of all squared loadings multiplying that factor. What is the amount of variance (in the six observed variables) that is explained by the first factor?

   The first factor explains $0.763^2 + 0.899^2 + 0.061^2 + 0.562^2 + 0.618^2 + 0.016^2 = 2.092$ of the total variance in these six variables.

   (d) What proportion of the total variance in the six observed variables is explained by the first factor?

   Since the total variance of the six (standardized) variables is 6, the percentage explained by the first factor is $2.092/6 = 34.86\%$.

   (e) Write a brief interpretation of the each factor. That is, characterize each factor if possible.

   Factor 1 is primarily related to symptoms and activity and to a lesser extent eating habits and appetite. We can see that as symptoms increase, so does daily activities and to a lesser degree eating habits and appetite. Factor 2 is primarily related to sleep and to a lesser extent eating habits and appetite. Here we can see that as sleep increases, so do eating habits and appetite. Factor 3 is primarily related to skin reaction. The factor suggests that skin reaction is not really related to any of the other variables.

2. For this problem use the "Pollution" data set. Columns correspond to wind, solar radiation, CO, NO, NO2, O3, and HC.

   (a) Do a factor analysis on all seven variables using the principal components method with two factors. Do a varimax rotation. Also, give the factor loadings after rotation. After rotation, the 2-factor loadings are

	Factor1	Factor2
wind	-0.12397	-0.47238
solarrad	-0.07574	0.69093
CO	0.70104	0.46719
NO	0.76309	-0.11300
NO2	0.76605	0.21919
O3	0.05060	0.83013
HC	0.60699	-0.03664

   Considering the variables are recorded on different scales (solarrad values are quite larger than the others), the standardized versions will be used throughout this part.

   (b) Write a brief characterization/interpretation of each factor.

   Factor 1 is primarily related to CO, NO, NO2 and HC. We can see that as CO increases, so do NO, NO2 and HC. Factor 2 is primarily a measure of O3. In addition, factor 2 is also considerably (but to a lesser extent) related to solarrad, wind, and CO. As O3, solarrad and CO increase, wind decreases.

   (c) Give the communalities (after rotation) for the analysis done in part a.

wind	solarrad	CO	NO	NO2	O3	HC
0.23851564	0.48311554	0.70973005	0.59508252	0.63488014	0.69168316	0.36977626

   (d) What is measured by the communalities given in part c?

   The communality estimates the amount of the variance of the seven variables that is explained by the two common factors. We can think of these values as multiple R^2 values for regression models predicting each of the seven variables from the two factors. The communality for each variable can be interpreted as the proportion of variation in that variable explained by the two factors.

   (e) Give the specific variance for each of the seven variables.

   By subtraction from one (since the variables were standardized), we find the specific variances as follows:

wind	solarrad	CO	NO	NO2	O3	HC
0.7614814	0.5168845	0.2902600	0.4049175	0.3651199	0.3083168	0.6302237

   (f) What proportion of the total variance of the seven variables is explained by the first factor?

   The total variance is 7, and the amount explained by each factor is given below. Thus, the fraction attributed to factor 1 is $2.0527/7 = 29.3\%$.

Factor1	Factor2
2.0527125	1.6700708

   (g) Repeat part a but with a three factor model.

   After rotation, the factor loadings are given below.

	Factor1	Factor2	Factor3
wind	-0.02870	-0.17361	0.84030
solarrad	0.04294	0.73598	-0.01674
CO	0.70551	0.27469	-0.38996
NO	0.64527	-0.38271	-0.48145
NO2	0.81137	0.15187	0.00409
O3	0.16599	0.81968	-0.15173
HC	0.70529	0.07134	0.46874

(h) Repeat part b for the three factor model.

Factor 1 is primarily related to NO2 and to a lesser degree to CO, HC, and NO. We can see that as CO increases, so do NO, NO2 and HC. Factor 2 is primarily related to O3 and solarrad. Here, we can see that as O3 increases, so does solarrad. Factor 3 is primarily a measure of wind. In addition, factor 3 is also considerably (but to a much lesser extent) related to NO and HC. As wind and HC increase, NO decreases.

(i) Repeat part c for the three factor model.

wind	solarrad	CO	NO	NO2	O3	HC
0.73707026	0.54378400	0.72527092	0.79463067	0.68139868	0.72244814	0.72224656

(j) Repeat part d for the three factor model.

The communality estimates the amount of the variance of the seven variables that is explained by the three common factors. We can think of these values as multiple R^2 values for regression models predicting each of the seven variables from the three factors. The communality for each variable can be interpreted as the proportion of variation in that variable explained by the three factors.

(k) Repeat part e for the three factor model.

wind	solarrad	CO	NO	NO2	O3	HC
0.2629297	0.4562160	0.2747291	0.2053603	0.3186013	0.2775519	0.2777534

(l) Repeat part f for the three factor model.

Factor1	Factor2	Factor3
2.1000843	1.4937579	1.3330071

Thus, the percent explained by the first factor is $2.1/7 = 30\%$.

SAS code:

```
data pollution;
  infile "v:\505\datasets\air.dat";
  input wind solarrad CO NO NO2 O3 HC;
  run;
proc factor data=pollution method=principal rotate=varimax nfactors=3;
  var wind solarrad CO NO NO2 O3 HC;
  run;
```