CS&SS Final Homework

Cameron Marsden

December 13, 2016

Problem 1

To examine the math data with latent class analysis, we should start by looking at the possible response patterns and their frequencies:

```
#Look at observed counts for each response pattern
resp.pat <- as.matrix(table(apply(math, 1, paste, collapse="")))
n.pat <- length(resp.pat)
poss.pat <- 2^8
perc.pat <- 100*(n.pat/poss.pat)
ordered.pat <- resp.pat[order(resp.pat,decreasing=TRUE),]
ordered.pat
```

```
00000000 11111111 00010000 11111101 11100001 11100000 11110111 00000100
     131       70       30       29       19       18       16       13
01111111 11110101 11111110 11111100 00010100 11000000 11101111 11110001
      11       11       10        8        7        7        7        7
01000000 11101101 11110100 00000001 00010001 00011100 01100000 10000000
       6        6        6        5        5        5        5        5
11011101 11011111 11101011 11110000 00011101 01110001 10100000 10111111
       5        5        5        4        3        3        3        3
11011110 11110110 11111011 00001100 00011111 01010000 01111101 01111110
       3        3        3        2        2        2        2        2
10110000 11010101 11011100 11100110 11101000 11101110 11111000 11111001
       2        2        2        2        2        2        2        2
00000101 00001000 00001101 00010101 00010111 00011001 01000001 01011110
       1        1        1        1        1        1        1        1
01011111 01100100 01100101 01101111 01110000 01110100 01110111 01111010
       1        1        1        1        1        1        1        1
01111011 10001001 10001101 10010000 10010101 10100001 11000001 11001111
       1        1        1        1        1        1        1        1
11010110 11010111 11011000 11011001 11011010 11100010 11100011 11100100
       1        1        1        1        1        1        1        1
11110011
       1
```

We have p=8 variables—leading to $2^8 = 256$ possible response patterns. Of these 256 patterns, only 81 (31.64) were actually observed. It's unsurprising that we will have many zero count cells given fraction type similarities across test questions. For example, Q4, Q10, Q11, and Q13 all present fractions with whole numbers while Q1-3 do not. So we would expect that if a student understands one fraction type, we would observe corresponding correct answers for all questions of that type.

Given that there are so many zero count cells, using the asymptotic χ^2 distribution for model selection will not yield reliable results. However, comparing these statistics between k vs. k+1 latent class models will give us good relative measures of overall fit.

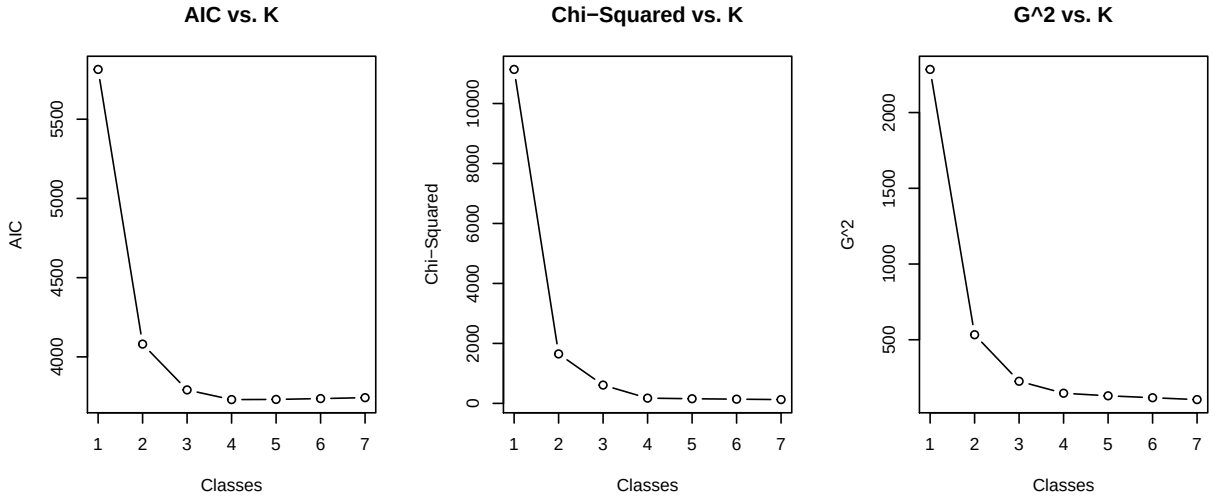
To determine how many latent classes are appropriate for the data, we run 50 repetitions of k=1,...,7 latent class models and gather the corresponding measures of fit reported in the table below:

```
kmods.tab <- rbind(AIC.k, df.k, chisq.k, Gsq.k,
                   1-pchisq(chisq.k,df.k),
                   1-pchisq(Gsq.k,df.k))
rownames(kmods.tab) <- c("AIC", "DF","Chi-Squared","G^2", "p-value Chi-Squared", "p-value G^2")
colnames(kmods.tab) <- c("k=1","k=2","k=3","k=4","k=5","k=6","k=7")
kable(kmods.tab,digits=4)
```

	k=1	k=2	k=3	k=4	k=5	k=6	k=7
AIC	5814.084	4080.683	3791.4661	3730.4276	3731.4778	3736.7455	3742.4937
DF	247.000	238.000	229.0000	220.0000	211.0000	202.0000	193.0000
Chi-Squared	11136.511	1649.980	610.5653	175.1966	155.4430	140.7525	126.9263
G^2	2284.647	533.246	226.0291	146.9906	130.0408	117.3085	105.0567
p-value Chi-Squared	0.000	0.000	0.0000	0.9884	0.9984	0.9997	0.9999
p-value G^2	0.000	0.000	0.5431	1.0000	1.0000	1.0000	1.0000

The following plots help us visualize these results:

```
#Plot the restults of each potential model
par(mfrow=c(1,3))
plot(AIC.k, main = "AIC vs. K",
     xlab = "Classes", ylab="AIC", type="b")
plot(chisq.k, main="Chi-Squared vs. K",
     xlab = "Classes", ylab="Chi-Squared",
     type="b")
plot(Gsq.k, main="G^2 vs. K",
     xlab = "Classes", ylab="G^2",
     type="b")
```



All three plots show an "elbow" at k=3 classes. Using AIC, χ^2, and G^2 as relative measure of fit for k=3 vs. k>3 provide evidence that using more than three classes is unnecessary. On the other hand, k<3 class models display noticeably poorer fit.

In addition, moving from three dimensions to four or more greatly reduces ease of interpretation.

Moving forward, we will analyze the three-latent class model

Problem 2

As mentioned previously, the stark differences in AIC, G^2 and χ^2 between k=2 and k=3 follwed by comparable values for each statistic between k=3 and k>3 provide evidence that k=3 is the most appropriate model of the ones we tested. However, analyzing the k=3 latent class model's fit on its own is more difficult. Using the asymptotic distribution χ^2 is inappropriate given large number of low-count cells.

One way to evaluate the chosen model is to collapse cells that are similar. Comparing all two and three-way marginal response patterns.

- Using the three-way marginal maximum, we know 100% of students answered Q1, Q2, Q3 either all correctly or all incorrectly.
- Of the remaining questions, we know 2300% of students answered Q10, Q11 either both correctly or both incorrectly.

Summing these two cases together, we can analyze the reduced five-column matrix to fit $k = 1,\ldots,4$ latent class models and compare with the full matrix results.

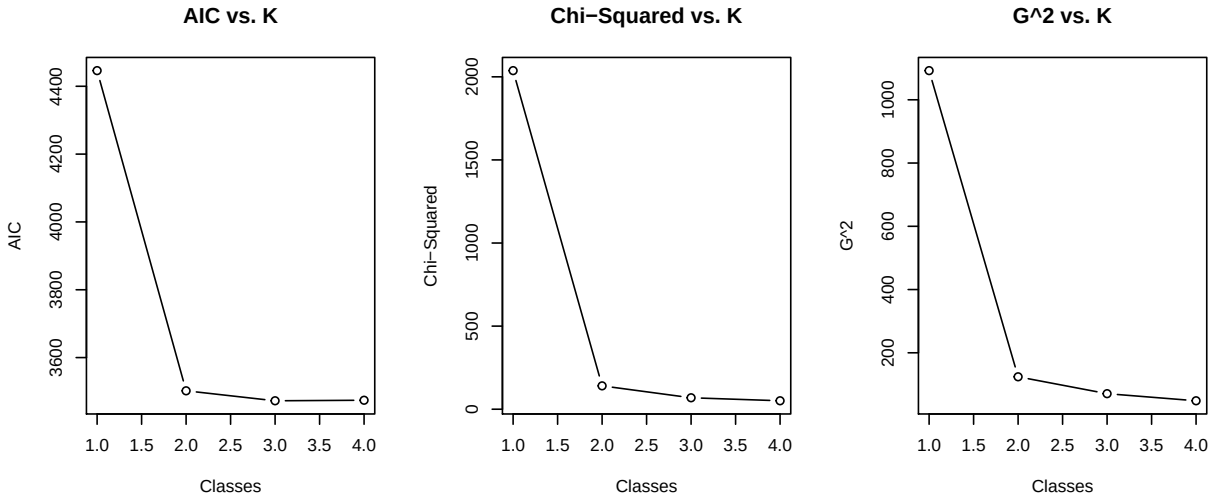
For the reduced columns dataset, there are $4 \times 3 \times 2^3 = 96$ possible cells, of which we observed 58 (60.42), almost twice the percentage of possible cells we observed for the full dataset.

Once again, we look at the summary statistics and plots and compare to those of the full data matrix:

```
kable(rkmods.tab,digits=4)
```

	k=1	k=2	k=3	k=4
AIC	4446.146	3502.0291	3472.6436	3474.2123
DF	180.000	168.0000	156.0000	144.0000
Chi-Squared	2037.242	140.5952	68.8255	50.8086
G^2	1092.224	124.1068	70.7213	48.2900
p-value Chi-Squared	0.000	0.9392	1.0000	1.0000
p-value G^2	0.000	0.9954	1.0000	1.0000

```
#Plot the restults of each potential model
par(mfrow=c(1,3))
plot(rAIC.k, main = "AIC vs. K",
     xlab = "Classes", ylab="AIC", type="b")
plot(rchisq.k, main="Chi-Squared vs. K",
     xlab = "Classes", ylab="Chi-Squared",
     type="b")
plot(rGsq.k, main="G^2 vs. K",
     xlab = "Classes", ylab="G^2",
     type="b")
```



Once again, the statistics suggest that one class is insufficient. When we collapse the data, two classes becomes more viable, but we still notice improvements in all three statistics for three-classes. Given the results from Problem 1, we will continue with the three-class model—being wary of how well the results support the model assumptions.

Problem 3

```
k <- 3 #number of classes
p <- 8 #number of variables
n <- length(math.mod[,1]) #sample size

#Extract each value of pi or 1-pi and eta
class.probs <- matrix(NA,ncol=p,nrow=k)
etas <- NULL
prob10 <- c(rep(2,3),rep(1,5))

for(i in 1:k){
  etas[i] <- mod.3$P[i]
  for(j in 1:p){
  class.probs[i,j] <- mod.3$probs[[j]][i,prob10[j]]
  }
}

#Multiply the probabilities for each class together with class proportion eta

class1 <- prod(class.probs[1,],etas[1])
class2 <- prod(class.probs[2,],etas[2])
class3 <- prod(class.probs[3,],etas[3])

#Sum the three segments together
sum.class <- sum(class1,class2,class3)

#Multiply by sample size
E.count <- n*sum.class
```

Under the three-class model, we find the probability of observing response pattern {1,1,1,0,0,0,0,0} as

$$\begin{aligned}
P(11100000) &= P(11100000|class1)P(class1) + P(11100000|class2)P(class2) + P(11100000|class3)P(class3) \\
&= \pi_{11}\pi_{21}\pi_{31}(1-\pi_{41})(1-\pi_{51})(1-\pi_{61})(1-\pi_{71})(1-\pi_{81})\eta_1 \\
&+ \pi_{12}\pi_{22}\pi_{32}(1-\pi_{42})(1-\pi_{52})(1-\pi_{62})(1-\pi_{72})(1-\pi_{82})\eta_2 \\
&+ \pi_{13}\pi_{23}\pi_{33}(1-\pi_{43})(1-\pi_{53})(1-\pi_{63})(1-\pi_{73})(1-\pi_{83})\eta_3 \\
&= 0.151 \prod 0.853, 0.914, 0.868, 0.763, 0.953, 0.948, 0.975, 0.594 \\
&+ 0.408 \prod 0.031, 0.036, 0, 0.755, 0.944, 0.854, 1, 0.922 \\
&+ 0.441 \prod 0.031, 0.036, 0, 0.755, 0.944, 0.854, 1, 0.922 \\
&= 0.04074 + 0 + 4 \times 10^{-5} \\
&= 0.04078
\end{aligned}$$

To get the expected count for {1,1,1,0,0,0,0,0}, we multiply the posterior probability and the sample size n:

$$\begin{aligned}
E(11100000) &= nP(11100000) \\
&= 21.85808
\end{aligned}$$

We can verify this result within R:

```
prob.R <- poLCA.predcell(mod.3,y=c(2,2,2,1,1,1,1,1))*n
E.count.R <- prob.R*n
```

$$P(1,1,1,0,0,0,0,0) = 21.858E[1,1,1,0,0,0,0,0] = 1.1715932 \times 10^4$$

```
obs.pat<-subset(math.mod,Q1==2& Q2==2& Q3==2& Q4==1& Q10==1& Q11==1& Q13==1& Q15==1)

obs.count <- length(obs.pat[,1])
diff.count <- obs.count-E.count.R
resid.count <- (diff.count^2)/E.count.R
```

In the data, we observed 18 students who matched this response pattern—giving us

$$Obs - Exp = -1.1697932 \times 10^4$$

and a standardized residual of

$$\frac{(Obs - Exp)^2}{Exp} = 1.167996 \times 10^4$$

The small residual indicates that the three-class model fits this particular response pattern well.

Looking at the questions themselves, we see that individuals with this response pattern correctly answered all questions involving common/improper fractions, but they also missed every single mixed fraction (i.e., with whole numbers).

Given that the model fits well for this pattern, we have a good estimate for the proportion of the student population who struggle with mixed fractions.

In addition, we could add weights to our model to give this pattern higher prevalence given its clear factor distinction: comprehension of fractions with whole numbers.

Problem 4

```
#p(11100000)
post.pat <- sum.class

#p(11100000|class)
p.pat.class1 <- prod(class.probs[1,])
p.class1 <- etas[1]
p.pat.class2 <- prod(class.probs[2,])
p.class2 <- etas[2]
p.pat.class3 <- prod(class.probs[3,])
p.class3 <- etas[3]

#p(class|11100000)
p.class1.pat <- (p.pat.class1*p.class1)/post.pat
p.class2.pat <- (p.pat.class2*p.class2)/post.pat
p.class3.pat <- (p.pat.class3*p.class3)/post.pat
```

To find the class assignment posterior probabilities for the same response pattern, {1,1,1,0,0,0,0,0}, we can use the results we found in Problem 3:

Class 1:

$$\begin{aligned} P(class1|11100000) &= \frac{P(class1, 11100000)}{P(11100000)} \\ &= \frac{P(11100000|class1)P(class1)}{P(11100000)} \\ &= \frac{P(11100000|class1)P(class1)}{P(11100000)} \\ &= \frac{0.151 \prod 0.853, 0.914, 0.868, 0.763, 0.953, 0.948, 0.975, 0.594}{0.041} \\ &= 0.99912 \end{aligned}$$

Class 2:

$$\begin{aligned} P(class2|11100000) &= \frac{P(class2, 11100000)}{P(11100000)} \\ &= \frac{P(11100000|class2)P(class2)}{P(11100000)} \\ &= \frac{P(11100000|class2)P(class2)}{P(11100000)} \\ &= \frac{0.408 \prod 0.031, 0.036, 0, 0.755, 0.944, 0.854, 1, 0.922}{0.041} \\ &= 0 \end{aligned}$$

Class 3:

$$\begin{aligned} P(class3|11100000) &= \frac{P(class3, 11100000)}{P(11100000)} \\ &= \frac{P(11100000|class3)P(class3)}{P(11100000)} \\ &= \frac{P(11100000|class3)P(class3)}{P(11100000)} \\ &= \frac{0.441 \prod 0.885, 0.962, 0.871, 0.11, 0.201, 0.078, 0.34, 0.187}{0.041} \\ &= 8.8 \times 10^{-4} \end{aligned}$$

We can compare these results with the R output:

```
est.assign <- poLCA.posterior(mod.3,y=c(2,2,2,1,1,1,1,1))
colnames(est.assign) <- c("P(class 1 | 11100000)",
                          "P(class 2 | 11100000)",
                          "P(class 3 | 11100000)")
rownames(est.assign) <- NULL
class.assigned <- which.max(est.assign)
class.post.prob <- max(est.assign)

kable(est.assign,digits=c(5,30,5))
```

P(class 1 | 11100000)	P(class 2 | 11100000)	P(class 3 | 11100000)
0.99912	0	0.00088

Our model assumes individuals belong to only one class. Once again, we see that the model fits this particular pattern very well as there is essentially no ambiguity that response pattern {1,1,1,0,0,0,0,0} falls under class 1 with probability

$$P(class1|11100000) = 0.99912$$

Problem 5

Overall, the three-latent class model fit the data quite well when looking at response pattern {1,1,1,0,0,0,0,0}, but given the large number of cells with low observed counts, we should look into the overall residuals of observed patterns.

In addition, we can look at the maximum posterior probability for each observed pattern. Given our model assumption of one class per pattern, we would hope to see the majority of maximum probabilities to be close to 1.

The table below reports important quantiles for the observed standardized residuals and maximum posterior probabilities for all observed patterns:

```
###Find summary stats on all observed and expected counts
all.obs <- ddply(math.mod,
                 ~Q1+Q2+Q3+Q4+Q10+Q11+Q13+Q15,
summarize,
n=length(Q15))

all.prob <- NULL
all.exp <- NULL
all.prob.poss <- NULL
all.exp.poss <- NULL
all.posteriors <- matrix(NA,
                         nrow=length(all.obs[,1]),
                         ncol=k)
all.max <- NULL

for(i in 1:length(all.obs[,1])){
  comp <- as.numeric(c(all.obs[i,c(1:8)]))
  all.prob[i] <- poLCA.predcell(mod.3,y=c(comp))
  all.posteriors[i,] <- poLCA.posterior(mod.3,y=c(comp))
  all.max[i] <- max(all.posteriors[i,])
  all.exp[i] <- all.prob[i]*n
}

all.diff<- all.obs$n-all.exp
all.resid <- (all.diff^2)/all.exp

resid.summ <- quantile(all.resid,
                       probs=c(0.05,0.1,0.25,
                               0.5,0.75,0.9,0.95))

max.summ <- quantile(all.max,probs=c(0.05,0.1,0.25,
                         0.5,0.75,0.9,0.95))

outliers.resid <- which(all.resid > 3)
n.outliers <- length(outliers.resid)
obs.outliers <- all.obs$n[outliers.resid]

outliers.1 <- length(which(obs.outliers==1))
max.outlier <- max(obs.outliers)

n.max.90 <- length(which(all.max>0.9))
n.max.99 <- length(which(all.max>0.99))
```

```
fit.summ <- rbind(resid.summ,max.summ)
colnames(fit.summ) <- c("5%","10%","25%","50%","75%","90%","95%")
rownames(fit.summ) <- c("Residuals","Maximum Posterior Probability")

kable(fit.summ,digits=4)
```

	5%	10%	25%	50%	75%	90%	95%
Residuals	0.0178	0.0507	0.2408	0.7366	2.3057	7.8955	35.0882
Maximum Posterior Probability	0.6333	0.7878	0.9221	0.9944	0.9996	1.0000	1.0000

Looking at the residuals summary, we can see that each observed pattern fits the model well through the third quartile. Of our 81 observed patterns, there are 18 residuals > 3. Unsurprisingly, these large residuals correspond to patterns that didn't appear often (10 only had one observed count, and the maximum residual had 7 observed counts). As mentioned in Problem 2, collapsing questions could help reduce the number of high residuals, improving the model fit.

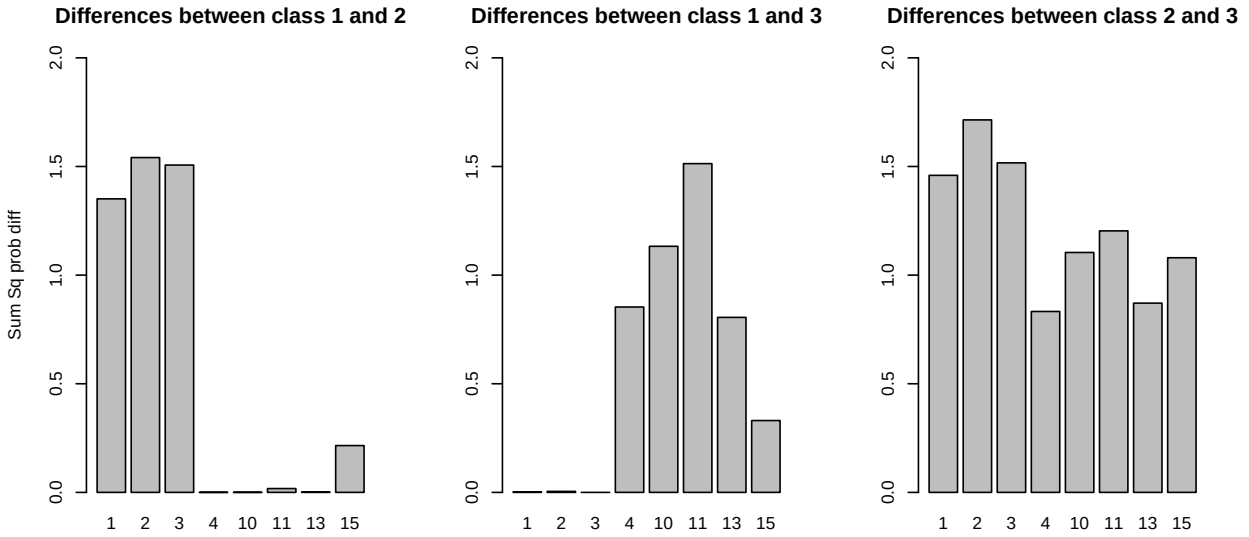
However, the summary of maximum residuals across all observed patterns is very optimistic as the vast majority of maximum posterior proportions for each unique pattern were very close to 1.

Given the number of empty/low count patterns, the AIC, χ^2, G^2, and posterior probabilities indicate that the three-class model fits well.

In order to interpret the latent classes themselves, it is easiest to observe the difference by looking at the squared differences in each pairwise $\pi_p j$ j=1,2,3 probabilities for each test question p=1,...,8. the barplots below show the three comparisons:

```
differences12 <- differences13 <- differences23 <- rep(0, 8)

for(j in 1:8){
differences12[j] <- sum((mod.3$probs[[j]][1, ] - mod.3$probs[[j]][2, ])^2)
}

for(j in 1:8){
differences13[j] <- sum((mod.3$probs[[j]][1, ] - mod.3$probs[[j]][3, ])^2)
}

for(j in 1:8){
differences23[j] <- sum((mod.3$probs[[j]][2, ] - mod.3$probs[[j]][3, ])^2)
}

par(mfrow=c(1,3))

barplot(differences12,
        names.arg = c(1,2,3,4,10,11,13,15),
main = "Differences between class 1 and 2",
ylab = "Sum Sq prob diff",ylim=c(0,2))

barplot(differences13,
        names.arg = c(1,2,3,4,10,11,13,15),
main = "Differences between class 1 and 3",
ylab = NULL,ylim=c(0,2))
```

```
barplot(differences23,
        names.arg = c(1,2,3,4,10,11,13,15),
main = "Differences between class 2 and 3",
ylab = NULL,ylim=c(0,2))
```



The first two plots show the clearest distinction between latenet classes. When comparing Class 1 and Class 3, we see the same pattern as the in-depth analysis of pattern {1,1,1,0,0,0,0,0}: Class 1 underlies students' ability to evaluate common/improper forms of fractions—with high probabilities indicating comprehension of this fractional form but not others. Class 3 forms the other side of the spectrum (i.e., students' ability to evaluate mixed fractions). High probabilities for class 3 indicate comprehension of whole number/mixed fractions at the expense of common/improper fraction comprehension.

The differences for Classes 2 and 3 aren't as highly delineated, however we can see higher values of our difference measure for the improper/common fraction questions. Considering that Class 3 showed a clear pattern of understanding mixed fractions but not improper/common, Class 2 most likely explains that students' with mixed understanding of the two types are more likely to have stronger comprehension of common/improper fractions. Looking at the column totals for each question corroborates this conclusion:

```
tot.col <- colSums(math)
tot.correct <- sum(math)
perc.col <- 100*tot.col/tot.correct
col.sum <- rbind(tot.col,round(perc.col,digits=2))
rownames(col.sum) <- c("# Correct by Question","% of Table Total")

Q123.mean <- round(mean(tot.col[1:3]),digits=2)
Qothers.mean <- round(mean(tot.col[4:8]),digits=2)

kable(col.sum)
```

	Q1	Q2	Q3	Q4	Q10	Q11	Q13	Q15
# Correct by Question	285.00	309.00	276.00	283.00	205.00	254.00	158.00	242.00
% of Table Total	14.17	15.36	13.72	14.07	10.19	12.62	7.85	12.03

Over all students and questions, there were 2012 correct answers given, which the table uses to show both the number of correct answers of each question and the weight (as percentage) of each question toward the total number of correct responses.

On average, Q1-3 gathered 290 total correct answers while the others (mixed fraction questions) gathered 228.4 correct answers.

Final Project Critque: *A Factor Analysis of Religious Regulation Laws amongst Monotheistic-Majority States* by Yusri Supiyan

Research Question and Data

Yusri's research focused on how heavily-monotheistic countries enforce constitutional restrictions based on faith—particularly for Islamic and Christian majority countries. His main goal was to determine any latent factors that could explain different methods of implementing restrictions among different religions.

Methods, Preliminary Analyses, Data Challenges and Approach

Yusri's research expounded on a study by Grim and Finke (2006) that explained how countries implement constitutional restrictions—namely with governmental regulation (e.g., laws) or religious favortism (e.g., governmental salaries for clergy). The dataset consisted of p=51 binary variables for n=177 countries. Each binary variable indicated whether or not a country implemented religion-based restrictions for a particular event (e.g., interfaith marriage, access to birth control, homosexuality). According to Yusri, the researchers observed each country's constitutions and de facto restrictions over the 19-year research period.

Yusri did not perform any preliminary analyses. I particularly asked about marginal table analysis and tetrachoric correlations. For performing the full analysis, he implemented a two-factor logistic factor model for all countries and a one-factor logistic model for both Muslim and Christian majority countries.

The biggest challenge for Yusri was the overwhelming presence of 0s (i.e., no restriction) within the data. To help simplify and bolster the results, he selected a subset of ten variables (out of the p=51 available) that he found particularly interesting and included less dominance of 0s.

Research Results

For the two-factor analysis on all countries, Yusri presented a bivariate plot of each country's loadings. He color coded each country's majority relgion (Christianity, Islam, Judaism, Other) to help decipher any noticeable patterns. This plot showed clear distinction of Islam vs. other religions for factor 1. Yusri explained this factor as the overall higher tendency to implement restrictions in Islamic states. The second factor didn't seem as clear, but Yusri explained it as the overall religious trend to implement restrictions on personal liberties dealing with gender, family, and sexual orientation.

Yusri plotted logistic curves for each one-factor analysis (Islam and Christianity, separately) to show the probability of countries implementing each of the ten restriction variables. The plots showed the trend of Islamic-majority ountries placing restrictions on nearly all variables while Christian-majority countries were likely to place restrictions only on the family/procreation variables.

I believe that Yusri provided great explanations of each factor model and his research question as a whole. The results made intuitive sense and simplified the trends of the ten religious restriction variables into one-two dimensional plots. However, Yusri didn't carry out any analysis of the model fit, so it was difficult to assess the validity of his findings. It also would have been useful if he clearly indicated how he expounded on Grim and Finke's research.

Other Possible Methods

Given Yusri's strictly binary variables, the only additional multivariate analysis method we covered in class that he could have applied is latent class analysis—which I think would be appropriate for his data given how similar it is to the math data used for the final homework. Approaching the research question with the same framework as the math data, Yusri would have more evidence about the differences in factors/classes between countries, and it could further support how many factors/classes are sufficient and interpretable.