Discrete Multivariate Analysis - 579 - HW #10

Alex Towell (atowell@siue.edu)

Supplemental material

We refactored the code for independence testing as a function that returns all relevant calculations given a matrix of cross-sectional data.

```
# csdata is cross sectional data (observed counts) entered as an IxJ matrix.
# tests for independence P(X,Y) = P(X)P(Y) using the X^2 statistic.
independence_test_X2 <- function(csdata)
{
  # define the dimensions of the table
  I = dim(csdata)[1]
  J = dim(csdata)[2]

  # compute the overall sample size, the row sample sizes, and the column sample sizes
  total = sum(csdata)
  row.sum = apply(csdata,1,sum)
  col.sum = apply(csdata,2,sum)

  # use matrix algebra to compute the table of expected cell counts
  expected = matrix(row.sum) %*% t(matrix(col.sum)) / total
  dimnames(expected) = dimnames(csdata)

  # compute the X^2 statistic, degrees of freedom, and p-value
  X2 = sum((obs-expected)^2/expected)
  df = (I-1)*(J-1)
  p.value.X2 = pchisq(X2,df,lower.tail = FALSE)

  # return computed values as a map
  list(expected_counts = expected,
       estimate = expected/sum(obs),
       X2 = X2,
       df = df,
       p.value = p.value.X2)
}
```

We also refactored the code for independence testing under the assumption of a monotonic association (γ correlation).

```
independence_test_gamma <- function(csdata)
{
  # to compute the estimate of gamma and the standard error, we need MESS
  library("MESS")

  # gkgamma takes an IxJ matrix as an input
  gk.result = gkgamma(csdata)
```

```
  # compute gamma.hat and its standard error
  gamma.hat = gk.result$estimate
  se = gk.result$se1

  # compute the z statistic and p-value for testing gamma=0 (independence)
  z = gamma.hat / se
  p.value.g = 2*pnorm(abs(z),lower.tail = FALSE)

  # display the results
  g.table = list(estimate=gamma.hat,
                 ase=se,
                 z.stat=z,
                 p.value=p.value.g)
}
```

Problem 1

Compute the statistic X^2 and the p-value for testing the model with independence between husband's rating and wife's rating. Provide an interpretation, stated in the context of the problem.

The observed data is cross-sectional with a general model given by

$$\{n_{ij}\} \sim \text{MULT}(n, \{\pi_{ij}\}).$$

We consider a simpler model where X and Y are independent, i.e., $\Pr(X, Y) = \Pr(X)\Pr(Y)$. Then, our task is to measure how compatible the observed data is to the null hypothesis

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}.$$

We prefer H_0 to a more general model that requires more parameters to estimate (and thus will have greater variance) and interpret.

We perform a chi-square test for independence. The test statistic is given by

$$X^2 = \sum_{i=1}^{4} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

where $\hat{m}_{ij} = n\hat{\pi}_{ij0}$ and $\hat{\pi}_{ij0} = \frac{n_{i+}n_{+j}}{n^2}$.

Under the null model, X^2 is distributed χ^2 with $\text{df} = 9$ degrees of freedom.

We compute the X^2 statistic and p-value with the following R code:

```
obs = matrix(c(7,7,2,3,2,8,3,7,1,5,4,9,2,8,9,14), nrow=4, byrow=TRUE)
colnames(obs) <- c("never/occassionally", "fairly often", "very often", "almost always")
rownames(obs) <- colnames(obs)

print(obs)
```

```
##                      never/occassionally fairly often very often almost always
## never/occassionally                   7            7          2             3
## fairly often                          2            8          3             7
## very often                            1            5          4             9
## almost always                         2            8          9            14
```

```
x2_test <- independence_test_X2(obs)
```

The observed X_0^2 is 16.955 and the p-value is 0.049. We consider this p-value to be moderate evidence against the null model. Stated differently, the observed data is moderately incompatible with the independence model.

For completeness, the estimates of π_{ij} under the independence model (with no additional assumptions) for the given cross-sectional data is given by:

```
round(x2_test$estimate,digits=3)
```

```
##                      never/occassionally fairly often very often almost always
## never/occassionally                0.028        0.064      0.041         0.076
## fairly often                       0.029        0.068      0.043         0.080
## very often                         0.028        0.064      0.041         0.076
## almost always                      0.048        0.112      0.072         0.132
```

Problem 2

Now compute the statistic Z^* and the p-value for testing the independence model using the correlation estimate $\hat{\gamma}$. Provide an interpretation, stated in the context of the problem.

The observed data is cross-sectional with a general model given by

$$\{n_{ij}\} \sim \text{MULT}(n, \{\pi_{ij}\}).$$

We consider a simpler model using γ,

$$H_0 : \gamma = 0,$$

i.e., γ is zero if X and Y are independent.

The test statistic is given by

$$Z^* = \frac{\hat{\gamma} - 0}{\hat{\sigma}(\hat{\gamma})},$$

which is normally distributed $\mathcal{N}(0, 1)$ under the null model.

We compute Z^* and p-value with the following R code:

```
gamma_test <- independence_test_gamma(obs)
```

The observed Z^* is 3.207 and the p-value is 0.001. We consider this p-value to be very strong evidence against the null model. That is, the observed data provides very strong evidence against the independence model when testing for support of a monotonic association model.

For completeness, the estimate for γ is $\hat{\gamma} = 0.36$.

Problem 3

Explain why a monotonic association model is better than a general association model in this example.

Estimates from simpler models have smaller variances than for estimates from more complicated models.