Rotating Principal Components: Diversity Data*

```
> rm(list=ls())
> # install.packages("readxl", dependencies = TRUE) # Only need to do this once
> library(readxl)
> # Download the data and put it in your working directory.
> # https://www.utstat.toronto.edu/brunner/data/legal/DiversityExplore.xlsx
> # The replication data are in DiversityReplic.xlsx
> ddata = read_excel("DiversityExplore.xlsx") # Read local copy

> ddata = as.data.frame(ddata) # Instead of a "tibble"
> dim(ddata); head(ddata)

[1] 500  47

  id Com1 Com2 Com3 Com4 Com5 Com6 Com7 Com8 Com9 Com10 RelC1 RelC2 RelC3 RelC4 RelC5
1  1    4    4    5    3    4    2    3    3    2     3     4     4     4     2     4
2  2    5    5    5    5    5    4    5    5    4     5     5     5     5     5     5
3  3    5    5    5    5    5    5    5    5    5     5     5     5     5     5     5
4  4    4    4    4    4    4    4    4    5    2     4     3     2     2     2     1
5  5    5    4    4    3    3    3    5    5    4     4     3     4     4     4     5
6  6    2    4    4    4    2    1    3    4    3     4     5     5     5     4     5
  RelM1 RelM2 RelM3 RelM4 RelM5 RelM6 RelM7 RelM8 RelM9 RelM10 RelM11 RelM12 Fair1
1     2     5     4     5     5     5     5     4     4      4      5      5     4
2     3     5     5     5     5     5     3     4     4      5      5      5     5
3     4     5     3     5     3     2     3     5     5      5      5      5     5
4     1     2     2     3     3     1     1     2     3      3      3      3     3
5     3     4     4     4     4     4     3     3     5      5      5      5     4
6     3     3     3     3     4     2     3     3     3      2      4      4     4
  Fair2 Fair3 Fair4 Fair5 Fair6 Sat1 Sat2 Sat3 Sat4 SM1 SM2 SM3 Gender VisMinority
1     1     4     1     2     1    2    2    3    4   3   1   2      0           1
2     4     5     5     5     5    5    4    5    5   4   5   5      1        <NA>
3     2     5     1     5     2    5    4    4    1   6   6   3      1           0
4     3     3     1     4     3    2    3    3    3   6   6   6      1           0
5     5     4     2     3     2    4    5    5    5   3   3   3      1           1
6     4     4     3     2     3    4    4    4    4   4   3   3      0           1
  EDUCLevel MaritalStatus Age CAN_Foreign_Born
1         4             2  28                0
2         4             2  41                0
3         5             3  45                0
4         3             2  39                0
5         7             1  26                0
6         3             2  59                0
```

*This handout was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The OpenOffice.org document is available from the course website:

http://www.utstat.toronto.edu/brunner/oldclass/431s23

```
> quest = as.matrix(ddata[,2:41])

> pc7 = prcomp(quest, scale = T, rank = 7) # Retain seven principal components
> ls(pc7)
 [1] "center"   "rotation" "scale"    "sdev"     "x"
```

$$
\begin{aligned}
\mathbf{z} &= \mathbf{L}\mathbf{f} + \mathbf{e} \\
&= \mathbf{L}\mathbf{R}^{\top}\mathbf{R}\mathbf{f} + \mathbf{e} \\
&= (\mathbf{L}\mathbf{R}^{\top})(\mathbf{R}\mathbf{f}) + \mathbf{e} \\
&= \mathbf{L}_2\mathbf{f}' + \mathbf{e}
\end{aligned}
$$

```
> # The seven principal components are in pc7$x
> L = cor(quest,pc7$x) # It really should be Lhat
> # round(L,4)
> vm7 = varimax(L); L2 = vm7$loadings
```

```
> print(L2,cutoff=0.3)

Loadings:
       PC1    PC2    PC3    PC4    PC5    PC6    PC7
Com1          -0.770
Com2          -0.564                              -0.379
Com3          -0.485                              -0.474
Com4          -0.710
Com5          -0.793
Com6          -0.772
Com7          -0.778
Com8          -0.752
Com9          -0.549
Com10         -0.423                              -0.591
RelC1                 0.796
RelC2                 0.812
RelC3                 0.725
RelC4                 0.702
RelC5  0.309          0.600                       -0.315
RelM1  0.686                                0.331
RelM2  0.767
RelM3  0.799
RelM4  0.819
RelM5  0.719
RelM6  0.766
RelM7  0.717
RelM8  0.719
RelM9  0.767
RelM10 0.726
RelM11 0.738
RelM12 0.762
Fair1  0.430                                0.425
Fair2  0.306 -0.334                         0.574
Fair3  0.310                                0.527
Fair4                                       0.529
Fair5                                       0.688
Fair6  0.351                                0.635
Sat1                            -0.674
Sat2                            -0.787
Sat3                            -0.816
Sat4                            -0.650
SM1                    -0.869
SM2                    -0.818
SM3                    -0.815

                 PC1   PC2   PC3   PC4   PC5   PC6   PC7
SS loadings    8.200 5.447 3.363 2.390 2.957 2.835 1.279
Proportion Var 0.205 0.136 0.084 0.060 0.074 0.071 0.032
Cumulative Var 0.205 0.341 0.425 0.485 0.559 0.630 0.662

> sum(L2[,3]^2)    # Amount of variance explained by component three.
[1] 3.363317
> sum(L2[,3]^2)/40 # Proportion of variance explained by component three.
[1] 0.08408292
>
> # If we had the questionnaire, we might be able to understand PC7
> # It explains only 3% of the variance. Try dropping it.
```

```
> pc6 = prcomp(quest, scale = T, rank = 6)
> L = cor(quest,pc6$x)
> vm6 = varimax(L); L2 = vm6$loadings
> print(L2,cutoff=0.3)

Loadings:
       PC1    PC2    PC3    PC4    PC5    PC6
Com1          -0.788
Com2          -0.642
Com3          -0.598
Com4          -0.736
Com5          -0.797
Com6          -0.764
Com7          -0.761
Com8          -0.688
Com9          -0.472                       0.318
Com10         -0.562
RelC1                 0.784
RelC2                 0.800
RelC3                 0.727
RelC4                 0.720
RelC5   0.316         0.628
RelM1   0.697
RelM2   0.772
RelM3   0.798
RelM4   0.819
RelM5   0.722
RelM6   0.771
RelM7   0.723
RelM8   0.715
RelM9   0.759
RelM10  0.727
RelM11  0.732
RelM12  0.754
Fair1   0.438                              0.409
Fair2   0.313                              0.603
Fair3   0.323         0.303                0.477
Fair4                                      0.557
Fair5                                      0.649
Fair6   0.361                              0.640
Sat1                                -0.658
Sat2                                -0.780
Sat3                                -0.812
Sat4                                -0.662
SM1                          -0.867
SM2                          -0.817
SM3                          -0.813

                 PC1   PC2   PC3   PC4   PC5   PC6
SS loadings    8.238 5.652 3.434 2.356 2.903 2.880
Proportion Var 0.206 0.141 0.086 0.059 0.073 0.072
Cumulative Var 0.206 0.347 0.433 0.492 0.565 0.637
```

```
> # It's always okay to reverse signs in a column.
> # It just reverses the meaning of the component.
> # Reverse columns 2, 4 and 5, and then name the components.
> L2[,c(2,4,5)] = -L2[,c(2,4,5)]
> colnames(L2) = c("RelMan", "CommitOrg", "RelCol", "SMCD", "JobSat", "FairOp")
> print(L2,cutoff=0.3)

Loadings:
       RelMan CommitOrg RelCol SMCD   JobSat FairOp
Com1           0.788
Com2           0.642
Com3           0.598
Com4           0.736
Com5           0.797
Com6           0.764
Com7           0.761
Com8           0.688
Com9           0.472                          0.318
Com10          0.562
RelC1                   0.784
RelC2                   0.800
RelC3                   0.727
RelC4                   0.720
RelC5   0.316           0.628
RelM1   0.697
RelM2   0.772
RelM3   0.798
RelM4   0.819
RelM5   0.722
RelM6   0.771
RelM7   0.723
RelM8   0.715
RelM9   0.759
RelM10  0.727
RelM11  0.732
RelM12  0.754
Fair1   0.438                                  0.409
Fair2   0.313                                  0.603
Fair3   0.323           0.303                  0.477
Fair4                                          0.557
Fair5                                          0.649
Fair6   0.361                                  0.640
Sat1                                   0.658
Sat2                                   0.780
Sat3                                   0.812
Sat4                                   0.662
SM1                            0.867
SM2                            0.817
SM3                            0.813

               RelMan CommitOrg RelCol  SMCD JobSat FairOp
SS loadings     8.238     5.652  3.434 2.356  2.903  2.880
Proportion Var  0.206     0.141  0.086 0.059  0.073  0.072
Cumulative Var  0.206     0.347  0.433 0.492  0.565  0.637
```

```
> # Produce rotated components for future use
```

$$
\begin{aligned}
\mathbf{z} &= \mathbf{L}\mathbf{f} + \mathbf{e} \\
&= \mathbf{L}\,\mathbf{R}^{\top}\mathbf{R}\,\mathbf{f} + \mathbf{e} \\
&= (\mathbf{L}\,\mathbf{R}^{\top})(\mathbf{R}\,\mathbf{f}) + \mathbf{e} \\
&= \mathbf{L}_2\mathbf{f}' + \mathbf{e}
\end{aligned}
$$

```
> n = dim(quest)[1]
> f = scale(pc6$x) * sqrt(n/(n-1)) # Divide by n, not n-1
> fprime = f %*% vm6$rotmat
> fprime[,c(2,4,5)] = -fprime[,c(2,4,5)]
> colnames(fprime) = colnames(L2)
> # Verify L2 = cor(data, fprime)
> round(cor(quest,fprime),3)
```

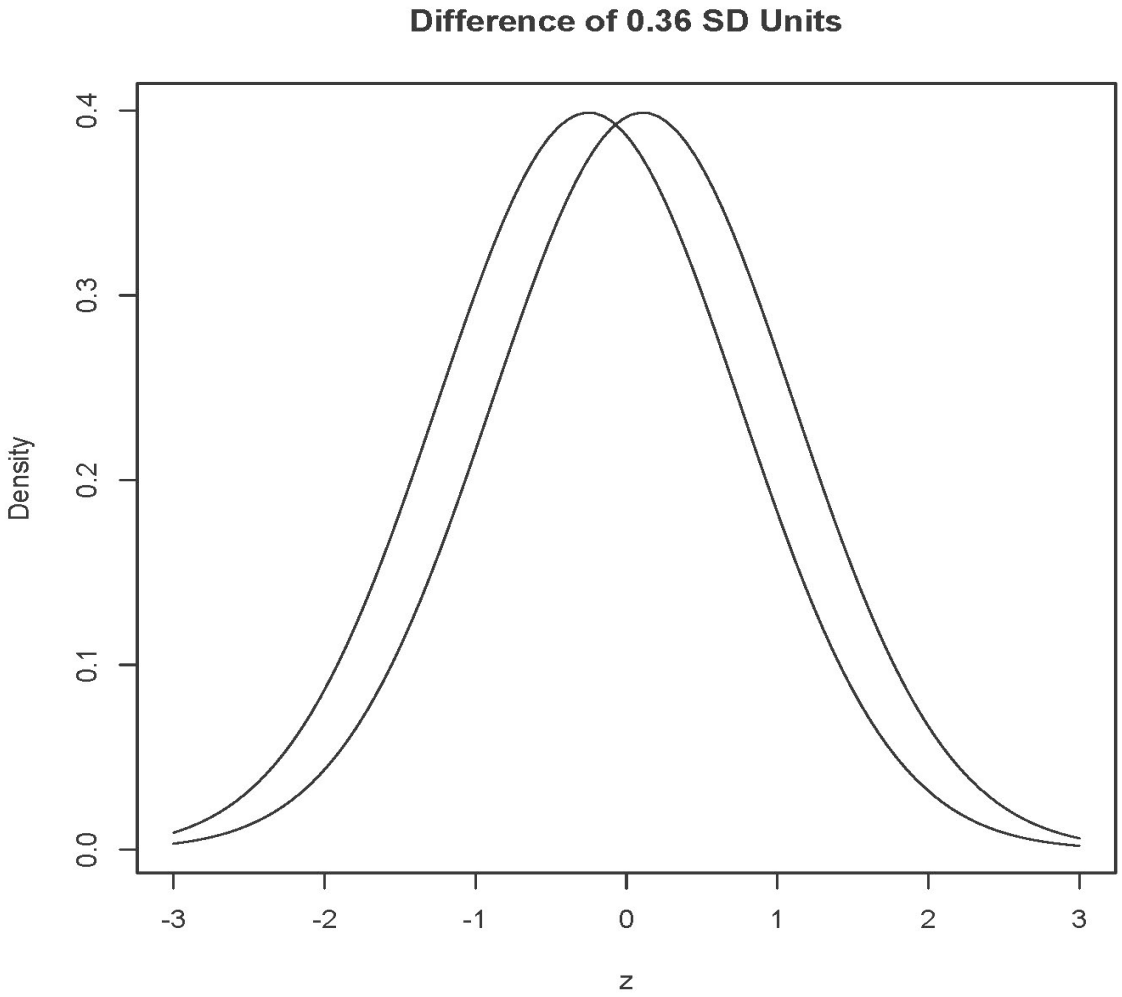
	RelMan	CommitOrg	RelCol	SMCD	JobSat	FairOp
Com1	0.193	0.788	0.090	0.015	0.078	0.172
Com2	0.093	0.642	0.219	0.070	0.117	0.063
Com3	0.136	0.598	0.088	0.046	0.197	-0.176
Com4	0.245	0.736	0.090	0.115	0.162	0.116
Com5	0.194	0.797	0.083	0.041	0.083	0.219
Com6	0.191	0.764	0.029	0.041	0.141	0.240
Com7	0.177	0.761	0.087	0.012	0.115	0.194
Com8	0.148	0.688	0.093	0.062	0.052	0.185
Com9	0.054	0.472	0.169	0.188	-0.070	0.318
Com10	0.066	0.562	0.121	0.123	0.039	-0.110
RelC1	0.148	0.174	0.784	0.110	0.103	0.083
RelC2	0.199	0.168	0.800	0.082	0.123	0.070
RelC3	0.266	0.208	0.727	0.018	0.134	0.157
RelC4	0.259	0.116	0.720	0.047	0.078	0.218
RelC5	0.316	0.124	0.628	0.034	0.112	0.102
RelM1	0.697	0.095	0.120	0.064	0.066	0.271
RelM2	0.772	0.195	0.152	0.074	0.083	0.154
RelM3	0.798	0.165	0.136	0.022	0.093	0.140
RelM4	0.819	0.179	0.137	0.038	0.056	0.121
RelM5	0.722	0.201	0.134	0.161	0.121	0.028
RelM6	0.771	0.102	0.093	0.072	0.168	0.190
RelM7	0.723	0.110	0.104	0.045	0.231	0.176
RelM8	0.715	0.157	0.110	0.054	0.208	0.076
RelM9	0.759	0.170	0.197	0.118	0.198	0.136
RelM10	0.727	0.085	0.174	0.051	0.114	0.222
RelM11	0.732	0.212	0.293	0.091	0.048	0.002
RelM12	0.754	0.159	0.178	0.054	0.160	0.129
Fair1	0.438	0.297	0.166	0.161	0.277	0.409
Fair2	0.313	0.289	0.094	0.130	0.235	0.603
Fair3	0.323	0.192	0.303	0.102	0.198	0.477
Fair4	0.117	0.092	0.149	-0.011	0.209	0.557
Fair5	0.295	0.102	0.109	0.043	0.134	0.649
Fair6	0.361	0.255	0.143	0.099	0.264	0.640
Sat1	0.220	0.223	0.089	0.010	0.658	0.211
Sat2	0.224	0.188	0.162	0.101	0.780	0.272
Sat3	0.275	0.143	0.134	0.080	0.812	0.241
Sat4	0.269	0.159	0.181	0.136	0.662	0.110
SM1	0.147	0.146	0.038	0.867	0.058	0.119
SM2	0.178	0.153	0.088	0.817	0.099	0.158
SM3	0.079	0.100	0.100	0.813	0.086	-0.048

```
> # Now use the components
> data.frame(1:47,colnames(ddata))
   X1.47  colnames.ddata.
1      1               id
2      2             Com1
3      3             Com2

      . . . skipping . . .

37    37             Sat3
38    38             Sat4
39    39              SM1
40    40              SM2
41    41              SM3
42    42           Gender
43    43      VisMinority
44    44        EDUCLevel
45    45    MaritalStatus
46    46              Age
47    47 CAN_Foreign_Born
>
> reduced_data = cbind(ddata[,42:47], fprime)
> head(reduced_data)
  Gender VisMinority EDUCLevel MaritalStatus Age CAN_Foreign_Born     RelMan
1      0           1         4             2  28                0  1.8068696
2      1        <NA>         4             2  41                0  0.3002811
3      1           0         5             3  45                0  0.1593790
4      1           0         3             2  39                0 -1.3827519
5      1           1         7             1  26                0  0.4704434
6      0           1         3             2  59                0 -0.9316213
   CommitOrg     RelCol        SMCD     JobSat      FairOp
1 -0.9318276 -0.7444822 -1.55046899 -0.3025774 -2.51396560
2  0.4817991  0.7869411  0.05914264  0.5861568  1.13900165
3  1.1006597  1.0363869  0.53291868 -1.0883484 -0.11154731
4  0.2049841 -3.3684570  2.02910905 -0.4534285  0.50342533
5 -0.4125786 -0.6716648 -0.92451060  1.2284831 -0.06030395
6 -1.7820171  1.5314054 -0.21785177  1.0008344 -0.16385212

> summary(reduced_data)

    Gender          VisMinority         EDUCLevel         MaritalStatus
 Length:500         Length:500         Length:500         Length:500
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


     Age            CAN_Foreign_Born       RelMan           CommitOrg
 Length:500         Length:500         Min.   :-3.4384   Min.   :-5.7629
 Class :character   Class :character   1st Qu.:-0.5859   1st Qu.:-0.4708
 Mode  :character   Mode  :character   Median : 0.2060   Median : 0.2068
                                       Mean   : 0.0000   Mean   : 0.0000
                                       3rd Qu.: 0.7559   3rd Qu.: 0.6645
                                       Max.   : 2.8086   Max.   : 2.0686
     RelCol             SMCD              JobSat            FairOp
 Min.   :-4.5656   Min.   :-2.657226   Min.   :-3.3066   Min.   :-2.9225
 1st Qu.:-0.5247   1st Qu.:-0.698009   1st Qu.:-0.6398   1st Qu.:-0.6407
 Median : 0.1219   Median :-0.007632   Median : 0.1016   Median : 0.1051
 Mean   : 0.0000   Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.6706   3rd Qu.: 0.659785   3rd Qu.: 0.7310   3rd Qu.: 0.7119
 Max.   : 2.5210   Max.   : 2.335664   Max.   : 2.6186   Max.   : 3.5956
```

```
> # The demographic variables are character. Fix.
> # Mostly make variables numeric indicators
> reduced_data = within(reduced_data,
+     {
+     Gender = as.numeric(Gender)
+     VisMinority  = as.numeric(VisMinority)
+     EDUCLevel  = as.numeric(EDUCLevel)
+     # Make MaritalStatus a factor, married = 2 is reference category
+     MaritalStatus  = factor(MaritalStatus)
+     contrasts(MaritalStatus) = contr.treatment(4, base = 2)
+     Age  = as.numeric(Age)
+     CAN_Foreign_Born = as.numeric(CAN_Foreign_Born)
+     })
>
> table(reduced_data$MaritalStatus, useNA = "ifany")

   1    2    3    4 <NA>
  76  383   34    4    3

> # Codes are 1 = Never married, 2 = Married, 3 = Divorced or separated,
> #           4 = Widowed
>
> head(reduced_data)

  Gender VisMinority EDUCLevel MaritalStatus Age CAN_Foreign_Born     RelMan
1      0           1         4             2  28                0  1.8068696
2      1          NA         4             2  41                0  0.3002811
3      1           0         5             3  45                0  0.1593790
4      1           0         3             2  39                0 -1.3827519
5      1           1         7             1  26                0  0.4704434
6      0           1         3             2  59                0 -0.9316213
   CommitOrg     RelCol        SMCD     JobSat      FairOp
1 -0.9318276 -0.7444822 -1.55046899 -0.3025774 -2.51396560
2  0.4817991  0.7869411  0.05914264  0.5861568  1.13900165
3  1.1006597  1.0363869  0.53291868 -1.0883484 -0.11154731
4  0.2049841 -3.3684570  2.02910905 -0.4534285  0.50342533
5 -0.4125786 -0.6716648 -0.92451060  1.2284831 -0.06030395
6 -1.7820171  1.5314054 -0.21785177  1.0008344 -0.16385212

>
> # Quick look at rotated PCs by binary demographics
> democor = cor(reduced_data[,c(1:3,5,6)], reduced_data[,7:12], use =
"pairwise.complete.obs")
> round(democor,3)

                RelMan CommitOrg RelCol   SMCD JobSat FairOp
Gender           0.012     0.009  0.015  0.057  0.036 -0.047
VisMinority      0.090    -0.009  0.035 -0.159 -0.166 -0.137
EDUCLevel       -0.030     0.007 -0.004 -0.019  0.023 -0.009
Age             -0.076     0.085  0.082 -0.001  0.026  0.017
CAN_Foreign_Born 0.038    -0.004 -0.029 -0.142 -0.152 -0.174
```

```
> # Eliminate all data with any missing values.
> noNAs = na.omit(reduced_data); dim(noNAs)
[1] 438  12
> r = cor( noNAs[,c(1:3,5,6)], noNAs[,7:12] )
> round(r , 3)
                 RelMan CommitOrg RelCol   SMCD JobSat FairOp
Gender            0.014    -0.005 -0.021  0.032  0.015 -0.028
VisMinority       0.090    -0.026  0.043 -0.156 -0.168 -0.122
EDUCLevel        -0.039    -0.016 -0.021 -0.022  0.000 -0.016
Age              -0.088     0.084  0.083 -0.003  0.018 -0.008
CAN_Foreign_Born  0.027     0.002  0.001 -0.126 -0.159 -0.150

>
> # Test: Are those correlations non-zero?
> n = dim(noNAs)[1]
> ttable = r * sqrt(n-2) / sqrt(1-r^2)
> round(ttable,3)

                 RelMan CommitOrg RelCol   SMCD JobSat FairOp
Gender            0.286    -0.100 -0.431  0.676  0.308 -0.593
VisMinority       1.887    -0.542  0.901 -3.295 -3.553 -2.561
EDUCLevel        -0.812    -0.327 -0.438 -0.456  0.009 -0.333
Age              -1.851     1.754  1.743 -0.054  0.378 -0.165
CAN_Foreign_Born  0.564     0.049  0.031 -2.655 -3.354 -3.159

> critval = qt(0.975,n-2); critval

[1] 1.96542

>
> # Bonferroni correcting for 30 tests: Divide alpha by 30
> a = 0.05/30
> Ccritval = qt(1-a/2,n-2); Ccritval

[1] 3.163715
```

```
> # Visible minority and job satisfaction: What is the actual mean difference?
> # Had r = -0.168, t = -3.553
> t.test(JobSat ~ VisMinority, var.equal = TRUE, data = reduced_data)

      Two Sample t-test

data:  JobSat by VisMinority
t = 3.6516, df = 469, p-value = 0.00029
alternative hypothesis: true difference in means between group 0 and group 1 is not equal
to 0
95 percent confidence interval:
 0.1656231 0.5515714
sample estimates:
mean in group 0 mean in group 1
      0.1079151      -0.2506822

>
> # What does a difference of 0.36 SDs look like?
> z = seq(-3,3,length = 100)
> Density = dnorm(z); d2 = dnorm(z, mean = 0.36)
> plot(z,Density, type = "l") # That's an L
> lines(z,d2,lty = 1); title("Difference of 0.36 SD Units")
```



```
> # One regression
> satisf = lm(JobSat ~ Gender + VisMinority + EDUCLevel +
+                       MaritalStatus + Age + CAN_Foreign_Born, data = reduced_data)
> summary(satisf)

Call:
lm(formula = JobSat ~ Gender + VisMinority + EDUCLevel + MaritalStatus +
    Age + CAN_Foreign_Born, data = reduced_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0896 -0.6056  0.1100  0.6893  3.0542

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.074023   0.308997   0.240   0.8108
Gender           0.058352   0.098358   0.593   0.5533
VisMinority     -0.246587   0.130071  -1.896   0.0587 .
EDUCLevel        0.012744   0.035297   0.361   0.7182
MaritalStatus1  -0.164540   0.148653  -1.107   0.2690
MaritalStatus3  -0.455440   0.188051  -2.422   0.0159 *
MaritalStatus4   0.098131   0.501503   0.196   0.8450
Age              0.001352   0.005782   0.234   0.8152
CAN_Foreign_Born -0.184632   0.131443  -1.405   0.1608
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9924 on 429 degrees of freedom
  (62 observations deleted due to missingness)
Multiple R-squared:  0.04955, Adjusted R-squared:  0.03183
F-statistic: 2.796 on 8 and 429 DF,  p-value: 0.005002
```

```
> # Well, one more
> sat2 = update(satisf, . ~ . - CAN_Foreign_Born) # Eliminate the Americans
> summary(sat2)

Call:
lm(formula = JobSat ~ Gender + VisMinority + EDUCLevel + MaritalStatus +
    Age, data = reduced_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3.06884 -0.65407  0.08495  0.69699  2.98565

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.050406   0.308888   0.163 0.870449
Gender         0.059055   0.098468   0.600 0.548999
VisMinority   -0.354023   0.105325  -3.361 0.000845 ***
EDUCLevel      0.015889   0.035266   0.451 0.652542
MaritalStatus1 -0.157541   0.148737  -1.059 0.290106
MaritalStatus3 -0.474241   0.187786  -2.525 0.011914 *
MaritalStatus4  0.124305   0.501724   0.248 0.804442
Age            0.001174   0.005787   0.203 0.839281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9936 on 430 degrees of freedom
  (62 observations deleted due to missingness)
Multiple R-squared:  0.04518, Adjusted R-squared:  0.02964
F-statistic: 2.907 on 7 and 430 DF,  p-value: 0.005581
```