

Week 7 Practical Session

David Barron

Trinity Term 2018

Factor analysis

You might like to explore the `psych` package, which has some additional features for exploratory factor analysis.

Let's have a look at the dataset `bfi`, which are responses to a personality test. There are supposed to be five factors: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. There are five items that are intended to load on each factor. More details can be found using `help('bfi')`.

Number of factors

There is no universally agreed upon way of deciding on the number of factors. In the `psych` package there is a function `VSS` that provides some tests.

```
library(psych)
data(bfi)
summary(bfi)
```

```
##           A1           A2           A3           A4
## Min.      :1.000   Min.      :1.000   Min.      :1.000   Min.      :1.0
## 1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.0
## Median :2.000   Median :5.000   Median :5.000   Median :5.0
## Mean     :2.413   Mean     :4.802   Mean     :4.604   Mean     :4.7
## 3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.0
## Max.     :6.000   Max.     :6.000   Max.     :6.000   Max.     :6.0
## NA's    :16     NA's    :27     NA's    :26     NA's    :19
##           A5           C1           C2           C3
## Min.      :1.00   Min.      :1.000   Min.      :1.00   Min.      :1.000
## 1st Qu.:4.00   1st Qu.:4.000   1st Qu.:4.00   1st Qu.:4.000
## Median :5.00   Median :5.000   Median :5.00   Median :5.000
## Mean     :4.56   Mean     :4.502   Mean     :4.37   Mean     :4.304
## 3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:5.000
## Max.     :6.00   Max.     :6.000   Max.     :6.00   Max.     :6.000
## NA's    :16     NA's    :21     NA's    :24     NA's    :20
##           C4           C5           E1           E2
## Min.      :1.000   Min.      :1.000   Min.      :1.000   Min.      :1.000
## 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
## Median :2.000   Median :3.000   Median :3.000   Median :3.000
## Mean     :2.553   Mean     :3.297   Mean     :2.974   Mean     :3.142
## 3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
## Max.     :6.000   Max.     :6.000   Max.     :6.000   Max.     :6.000
## NA's    :26     NA's    :16     NA's    :23     NA's    :16
##           E3           E4           E5           N1
## Min.      :1.000   Min.      :1.000   Min.      :1.000   Min.      :1.000
## 1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:2.000
## Median :4.000   Median :5.000   Median :5.000   Median :3.000
## Mean     :4.001   Mean     :4.422   Mean     :4.416   Mean     :2.929
## 3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:4.000
```

```

## Max. :6.000 Max. :6.000 Max. :6.000 Max. :6.000
## NA's :25 NA's :9 NA's :21 NA's :22
## N2 N3 N4 N5
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.00
## Median :4.000 Median :3.000 Median :3.000 Median :3.00
## Mean :3.508 Mean :3.217 Mean :3.186 Mean :2.97
## 3rd Qu.:5.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.00
## Max. :6.000 Max. :6.000 Max. :6.000 Max. :6.00
## NA's :21 NA's :11 NA's :36 NA's :29
## O1 O2 O3 O4
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000
## 1st Qu.:4.000 1st Qu.:1.000 1st Qu.:4.000 1st Qu.:4.000
## Median :5.000 Median :2.000 Median :5.000 Median :5.000
## Mean :4.816 Mean :2.713 Mean :4.438 Mean :4.892
## 3rd Qu.:6.000 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.:6.000
## Max. :6.000 Max. :6.000 Max. :6.000 Max. :6.000
## NA's :22 NA's :28 NA's :14
## O5 gender education age
## Min. :1.00 Min. :1.000 Min. :1.00 Min. : 3.00
## 1st Qu.:1.00 1st Qu.:1.000 1st Qu.:3.00 1st Qu.:20.00
## Median :2.00 Median :2.000 Median :3.00 Median :26.00
## Mean :2.49 Mean :1.672 Mean :3.19 Mean :28.78
## 3rd Qu.:3.00 3rd Qu.:2.000 3rd Qu.:4.00 3rd Qu.:35.00
## Max. :6.00 Max. :2.000 Max. :5.00 Max. :86.00
## NA's :20 NA's :223

```

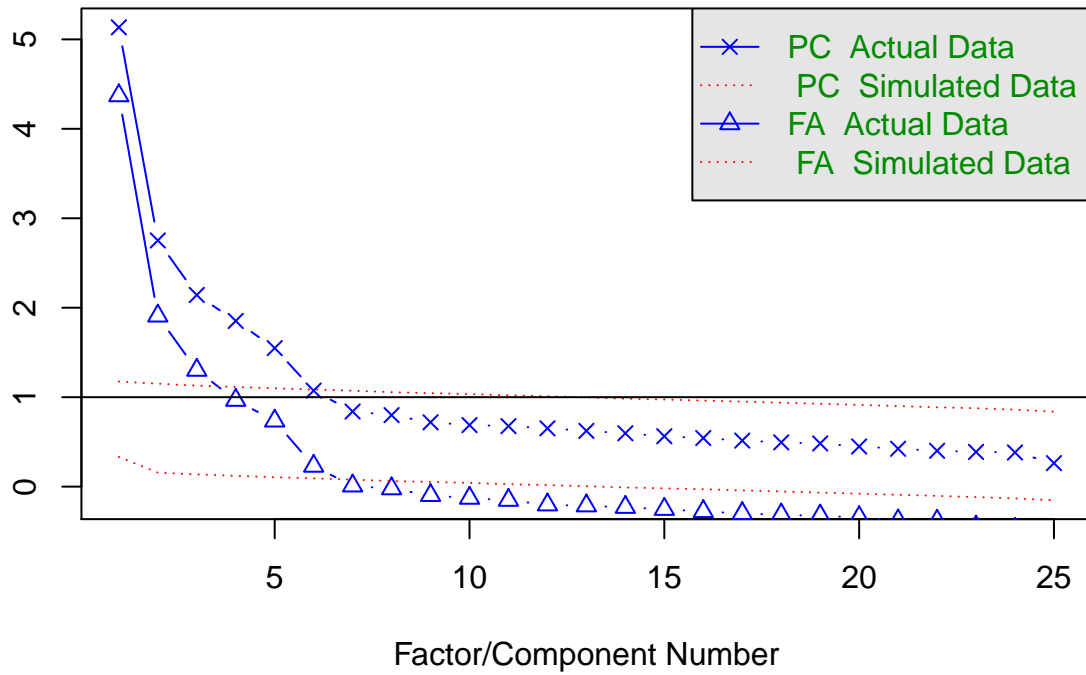
```

bfi.R <- cor(bfi[, 1:25], use = 'complete.obs')
p1 <- principal(bfi.R, rotate='varimax', n.obs = 2800)
fa.parallel(bfi.R, n.obs=2800, fm='ml')

```

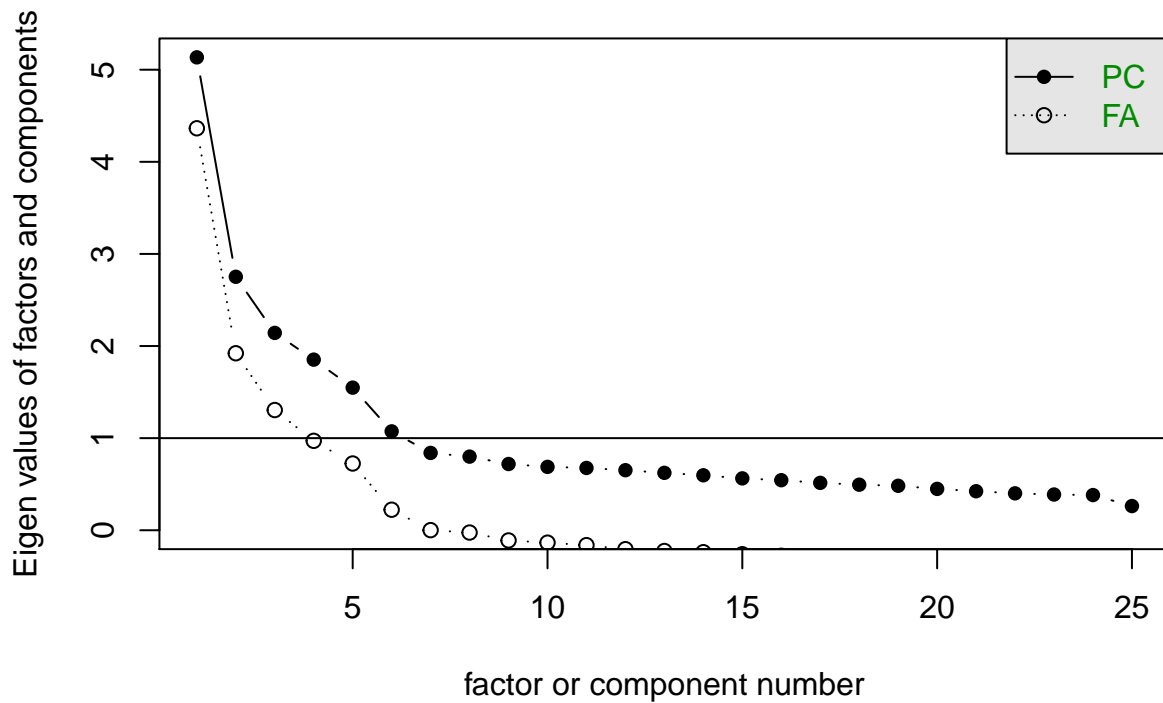
eigenvalues of principal components and factor analysis

Parallel Analysis Scree Plots



```
## Parallel analysis suggests that the number of factors = 6 and the number of components = 5
#v1 <- VSS(bfi.R, 8, fm='ml', n.obs = 2800)
#v1
scree(bfi.R)
```

Scree plot



It's not very clear whether you need four or five factors. Five is the number there are supposed to be (but there are lots of reasons why this might not work in practice). Let's compare the five and four factor solutions.

Five factors

```
ml5.out <- fa(bfi.R, nfactors = 5, rotate = "varimax", fm='ml', n.obs = 2800)
ml5.out$loadings
```

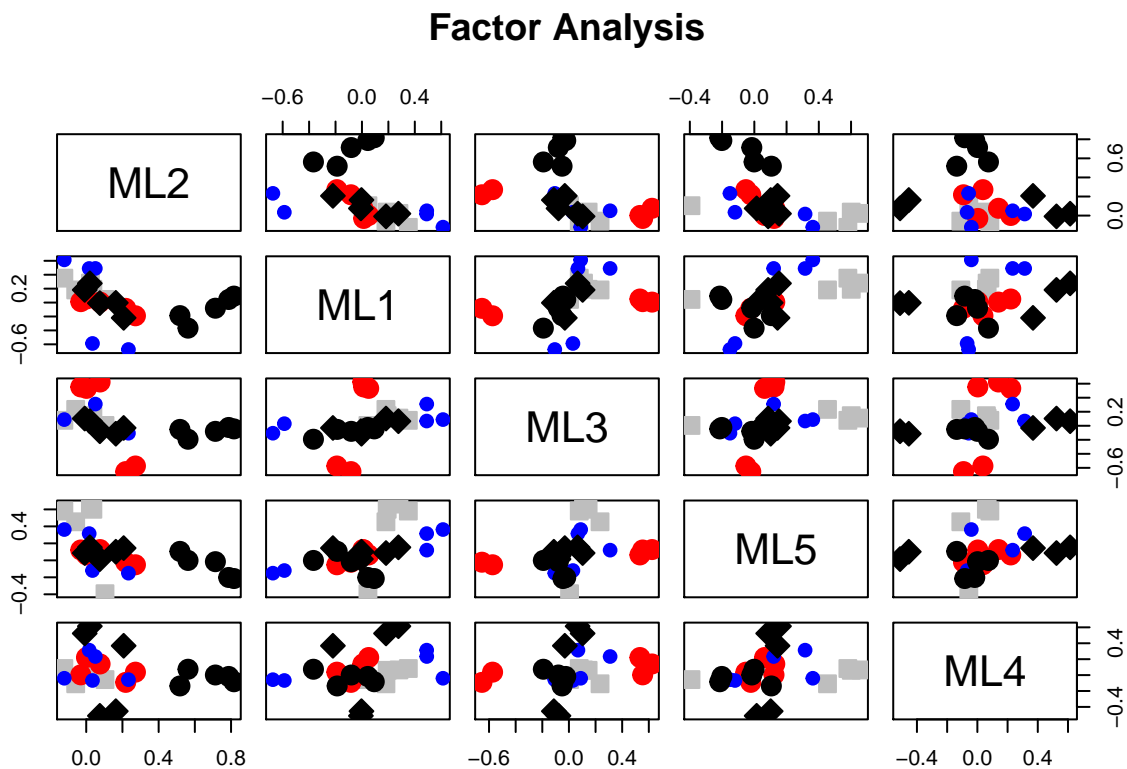
```
##
## Loadings:
##      ML2    ML1    ML3    ML5    ML4
## A1  0.104          -0.393
## A2          0.191  0.144  0.601
## A3          0.280  0.110  0.662
## A4          0.181  0.234  0.454 -0.109
## A5 -0.124  0.351          0.580
## C1          0.533          0.221
## C2          0.624  0.127  0.140
## C3          0.554  0.122
## C4  0.218          -0.653
## C5  0.272 -0.190 -0.573
## E1          -0.587          -0.120
## E2  0.233 -0.674 -0.106 -0.151
## E3          0.490          0.315  0.313
## E4 -0.121  0.613          0.363
```

```

## E5      0.491  0.310  0.120  0.234
## N1  0.816                -0.214
## N2  0.787                -0.202
## N3  0.714
## N4  0.562 -0.367 -0.192
## N5  0.518 -0.187          0.106 -0.137
## O1      0.182  0.103          0.524
## O2  0.163          -0.113  0.102 -0.454
## O3      0.276          0.153  0.614
## O4  0.207 -0.220          0.144  0.368
## O5                -0.512
##
##                ML2  ML1  ML3  ML5  ML4
## SS loadings    2.687 2.320 2.034 1.978 1.557
## Proportion Var 0.107 0.093 0.081 0.079 0.062
## Cumulative Var 0.107 0.200 0.282 0.361 0.423

```

```
plot(ml5.out, cut=0.3, cex = 2)
```



Looking at the loadings, you can see that the Openness item seems the least clear.

Four factors

```

ml4.out <- fa(bfi.R, nfactors = 4, rotate = "varimax", fm='ml', n.obs = 2800)
ml4.out$loadings

```

```

##
## Loadings:
##   ML1   ML2   ML3   ML4
## A1 -0.199  0.133
## A2  0.511         0.145
## A3  0.615         0.110
## A4  0.424         0.227 -0.138
## A5  0.639 -0.151
## C1         0.534  0.209
## C2  0.104         0.620  0.124
## C3         0.554
## C4         0.221 -0.658
## C5 -0.178  0.271 -0.566
## E1 -0.517
## E2 -0.592  0.222 -0.104
## E3  0.599         0.301
## E4  0.711 -0.121
## E5  0.461         0.299  0.240
## N1         0.810
## N2         0.793
## N3         0.719
## N4 -0.268  0.547 -0.181
## N5         0.510        -0.148
## O1  0.215         0.518
## O2         0.172 -0.114 -0.470
## O3  0.332         0.612
## O4         0.183         0.322
## O5         -0.517
##
##           ML1   ML2   ML3   ML4
## SS loadings  3.281 2.675 2.008 1.515
## Proportion Var 0.131 0.107 0.080 0.061
## Cumulative Var 0.131 0.238 0.319 0.379

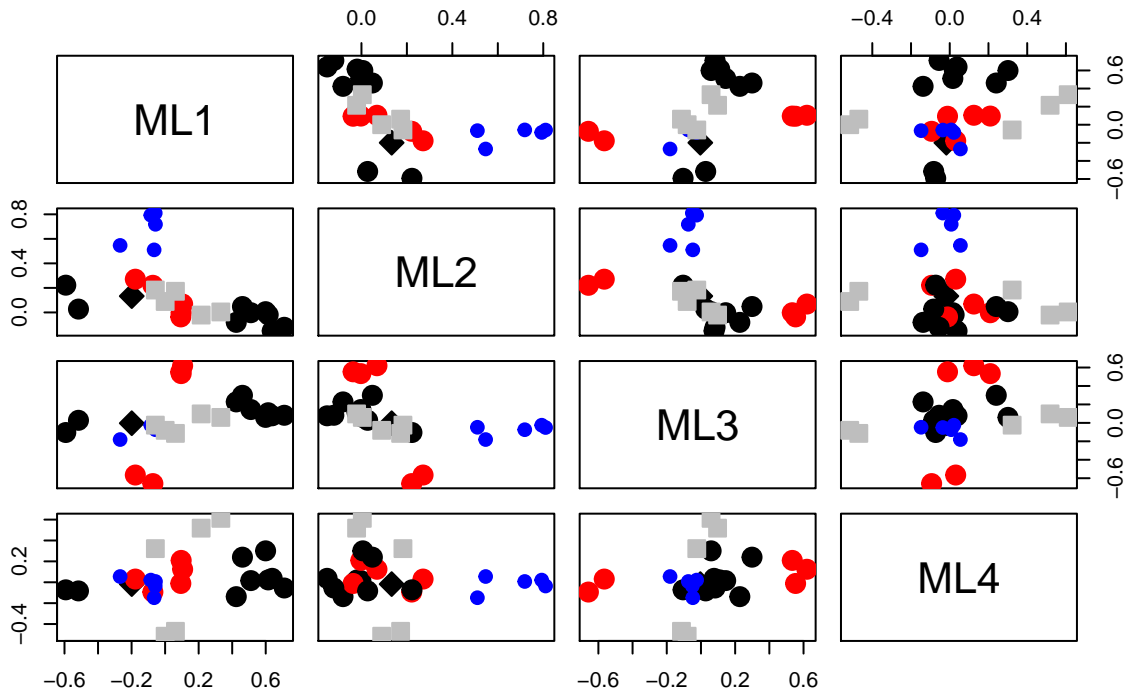
```

```

plot(ml4.out, cut=0.3, cex = 2)

```

Factor Analysis



This solution fails to distinguish between Agreeableness and Extraversion, so the five factor solution is probably to be preferred.

We can look to see if there are differences based on gender, education and age.

```
bfi.scores <- factor.scores(bfi[1:25], ml5.out)
head(bfi.scores$scores)
```

```
##           ML2           ML1           ML3           ML5           ML4
## 61617 -0.4354757  0.36611194 -1.33325905 -0.8899694 -1.8207464
## 61618  0.0633679  0.56445587 -0.70376255 -0.1175545 -0.1533549
## 61620  0.6395802  0.48282487  0.05552719 -0.9701237  0.2594642
## 61621 -0.1553552  0.07928884 -1.18765462  0.2455958 -1.0712305
## 61622 -0.4324621  0.54197994 -0.04594893 -0.8651299 -0.7552646
## 61623  0.3331240  1.22617711  1.52918815 -0.2171455  0.3692746
```

```
for (i in 1:5){
  print(t.test(bfi.scores$scores[, i], bfi$gender))
  print(summary(lm(bfi.scores$scores[, i] ~ age + education, data = bfi)))
}
```

```
##
## Welch Two Sample t-test
##
## data: bfi.scores$scores[, i] and bfi$gender
## t = -74.835, df = 3341.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```

## -1.707024 -1.619860
## sample estimates:
## mean of x mean of y
## 0.008343793 1.671785714
##
##
## Call:
## lm(formula = bfi.scores$scores[, i] ~ age + education, data = bfi)
##
## Residuals:
## Min 1Q Median 3Q Max
## -3.00410 -0.78312 -0.08146 0.71940 2.79000
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.269872 0.078364 3.444 0.000584 ***
## age -0.008189 0.002048 -3.999 6.56e-05 ***
## education -0.011131 0.019643 -0.567 0.570996
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9998 on 2233 degrees of freedom
## (564 observations deleted due to missingness)
## Multiple R-squared: 0.00826, Adjusted R-squared: 0.007371
## F-statistic: 9.299 on 2 and 2233 DF, p-value: 9.513e-05
##
##
## Welch Two Sample t-test
##
## data: bfi.scores$scores[, i] and bfi$gender
## t = -76.001, df = 3344.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.730773 -1.643718
## sample estimates:
## mean of x mean of y
## -0.01545947 1.67178571
##
##
## Call:
## lm(formula = bfi.scores$scores[, i] ~ age + education, data = bfi)
##
## Residuals:
## Min 1Q Median 3Q Max
## -3.6323 -0.6373 0.0664 0.7433 2.9443
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.028181 0.078375 0.360 0.7192
## age 0.002475 0.002048 1.209 0.2269
## education -0.034711 0.019645 -1.767 0.0774 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```



```

## Residual standard error: 0.9999 on 2233 degrees of freedom
## (564 observations deleted due to missingness)
## Multiple R-squared: 0.001677, Adjusted R-squared: 0.0007831
## F-statistic: 1.876 on 2 and 2233 DF, p-value: 0.1535
##
##
## Welch Two Sample t-test
##
## data: bfi.scores$scores[, i] and bfi$gender
## t = -74.8, df = 3333.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.713080 -1.625566
## sample estimates:
## mean of x mean of y
## 0.002462332 1.671785714
##
##
## Call:
## lm(formula = bfi.scores$scores[, i] ~ age + education, data = bfi)
##
## Residuals:
## Min 1Q Median 3Q Max
## -3.5472 -0.6555 0.0785 0.7557 2.3806
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.159147 0.078184 -2.036 0.0419 *
## age 0.007611 0.002043 3.725 0.0002 ***
## education -0.008840 0.019598 -0.451 0.6520
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9975 on 2233 degrees of freedom
## (564 observations deleted due to missingness)
## Multiple R-squared: 0.006284, Adjusted R-squared: 0.005394
## F-statistic: 7.06 on 2 and 2233 DF, p-value: 0.0008777
##
##
## Welch Two Sample t-test
##
## data: bfi.scores$scores[, i] and bfi$gender
## t = -74.86, df = 3325.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.721508 -1.633633
## sample estimates:
## mean of x mean of y
## -0.005784746 1.671785714
##
##
## Call:
## lm(formula = bfi.scores$scores[, i] ~ age + education, data = bfi)
##

```

```

## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.4183 -0.5714  0.1173  0.6996  2.5763
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.194969   0.077594  -2.513 0.012051 *
## age          0.006796   0.002028   3.352 0.000816 ***
## education    0.008162   0.019450   0.420 0.674778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.99 on 2233 degrees of freedom
## (564 observations deleted due to missingness)
## Multiple R-squared:  0.005749, Adjusted R-squared:  0.004859
## F-statistic: 6.456 on 2 and 2233 DF, p-value: 0.0016
##
## Welch Two Sample t-test
##
## data: bfi.scores$scores[, i] and bfi$gender
## t = -74.68, df = 3350.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.696244 -1.609455
## sample estimates:
## mean of x mean of y
## 0.01893589 1.67178571
##
## Call:
## lm(formula = bfi.scores$scores[, i] ~ age + education, data = bfi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5010 -0.6592  0.0374  0.7285  2.5794
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.357328   0.077861  -4.589 4.69e-06 ***
## age          0.001496   0.002034   0.735  0.462
## education    0.108372   0.019517   5.553 3.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9934 on 2233 degrees of freedom
## (564 observations deleted due to missingness)
## Multiple R-squared:  0.0157, Adjusted R-squared:  0.01482
## F-statistic: 17.81 on 2 and 2233 DF, p-value: 2.121e-08

```

There are significant gender differences in personality on all the dimensions. Some of them vary with age and education. It's plausible that education could be associated with personality (though you'd expect the causal direction to be the other way around), but not age!

Homework

1. Use the dataset `FactorAnalysis.csv` (which can be read using `read.csv()` or `read_csv`).
2. The dataset has 300 rows and 6 variables consisting of university students' ratings of their liking of six subjects on a five point scale from 1 = Strongly Dislike to 5 = Strongly Like. The six subjects are:
 - a. BIO (biology)
 - b. GEO (geology)
 - c. CHEM (chemistry)
 - d. ALG (algebra)
 - e. CALC (calculus)
 - f. STAT (statistics)
3. Conduct an exploratory factor analysis. How many factors are appropriate? How would you interpret these factors?