SEM 2019

Mikko Patronen

22 January, 2019

Week 1

Exercise 1.2

a) LINEAR REGRESSION

In this exercise a linear regression model was built for one continuous observed dependent variable (y1) with two covariates (x1 and x3). The data “ex3.1” was first manually imported into R and saved as .Rdata -file with R code lines:

df <- ex3.1

save(df, file=“df.Rdata”)

Here is a summary of the variables:

```
load("df.Rdata")
summary(df)
```

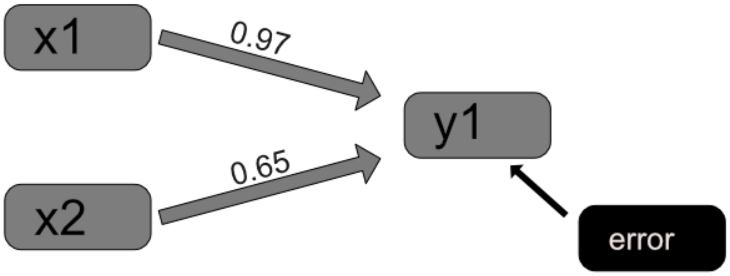
```
##        V1                 V2                  V3          
##  Min.   :-4.1163   Min.   :-3.145148   Min.   :-3.13875  
##  1st Qu.:-0.5269   1st Qu.:-0.749801   1st Qu.:-0.75466  
##  Median : 0.4288   Median : 0.023194   Median :-0.04029  
##  Mean   : 0.4848   Mean   : 0.001289   Mean   :-0.04216  
##  3rd Qu.: 1.5721   3rd Qu.: 0.755620   3rd Qu.: 0.71940  
##  Max.   : 5.1110   Max.   : 2.920440   Max.   : 2.87514  
```

Then a model was built according to instructions (y1 is the dependent variable, x1 and x3 are independent explanatory variables):

```
y1 <- df$V1
x1 <- df$V2
x3 <- df$V3

model <- lm(y1 ~ x1 + x3)
summary(model)
```

```
## 
## Call:
## lm(formula = y1 ~ x1 + x3)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.1506 -0.5752  0.0235  0.5663  3.1899 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.51096    0.04356   11.73   <2e-16 ***
## x1           0.96949    0.04163   23.29   <2e-16 ***
## x3           0.64904    0.04451   14.58   <2e-16 ***
```




```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9731 on 497 degrees of freedom
## Multiple R-squared:  0.609,  Adjusted R-squared:  0.6075
## F-statistic: 387.1 on 2 and 497 DF,  p-value: < 2.2e-16
```

According to the results both covariates x1 and x3 are statistically significant ($p < 0.001$). They both have a positive effect on the variable y1: when x1 increases one unit, the variable y1 increases 0.97 units (when x3 is considered a constant) and when x3 increases one unit, the variable y1 increases 0.65 units (when x1 is considered a constant). The model explains around 60% of the variance in the variable y1 (Adjusted R-squared = 0.6075).

The graph of the model is on top of this page (drawn with Affinity Designer):

b) EXPLORATORY FACTOR ANALYSIS

In this part an exploratory factor analysis is conducted according to instructions. The data file "ex4.1a" was imported manually into R and then wrangled so that the analysis could be run. The wrangling code is here:

df2 <- ex4.1a

colnames(df2) <- c("y1", "y2", "y3", "y4", "y5", "y6", "y7", "y8", "y9", "y10", "y11", "y12")

save(df2, file="df2.Rdata")

Let us view a summary of the data:

```r
load("df2.Rdata")
summary(df2)
```

```
##        y1                  y2                  y3
##  Min.   :-2.886760   Min.   :-3.69059   Min.   :-2.588919
##  1st Qu.:-0.682516   1st Qu.:-0.61723   1st Qu.:-0.673121
##  Median : 0.013133   Median : 0.06940   Median :-0.071101
##  Mean   : 0.008001   Mean   : 0.03339   Mean   : 0.003162
##  3rd Qu.: 0.700274   3rd Qu.: 0.69136   3rd Qu.: 0.689685
##  Max.   : 2.529128   Max.   : 2.79520   Max.   : 2.967696
##        y4                  y5                  y6
##  Min.   :-3.214602   Min.   :-2.94869   Min.   :-2.500254
##  1st Qu.:-0.577758   1st Qu.:-0.56400   1st Qu.:-0.630876
##  Median :-0.006558   Median : 0.04973   Median :-0.007958
##  Mean   : 0.073489   Mean   : 0.06330   Mean   : 0.062216
```

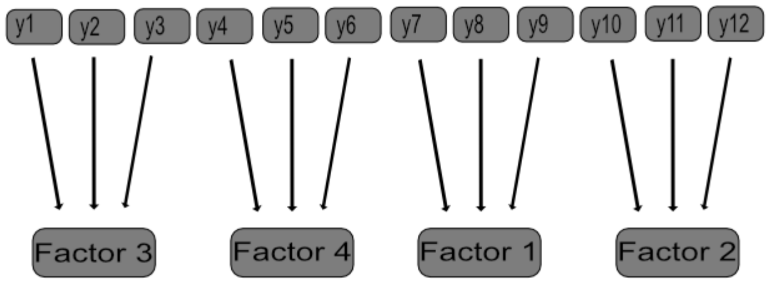
```
##  3rd Qu.: 0.768797   3rd Qu.: 0.76779   3rd Qu.: 0.792593
##  Max.   : 2.892782   Max.   : 3.74102   Max.   : 3.253644
##        y7                  y8                  y9
##  Min.   :-2.798568   Min.   :-3.581810   Min.   :-2.76235
##  1st Qu.:-0.631859   1st Qu.:-0.608176   1st Qu.:-0.64894
##  Median : 0.002374   Median : 0.030146   Median :-0.04405
##  Mean   :-0.003501   Mean   : 0.009048   Mean   : 0.02085
##  3rd Qu.: 0.688036   3rd Qu.: 0.692113   3rd Qu.: 0.69171
##  Max.   : 3.446497   Max.   : 2.827687   Max.   : 2.93974
##       y10                 y11                 y12
##  Min.   :-3.62913   Min.   :-2.747190   Min.   :-3.442931
##  1st Qu.:-0.75897   1st Qu.:-0.680559   1st Qu.:-0.706488
##  Median : 0.01185   Median : 0.024163   Median :-0.008250
##  Mean   :-0.03686   Mean   : 0.001595   Mean   :-0.002375
##  3rd Qu.: 0.63595   3rd Qu.: 0.692018   3rd Qu.: 0.655408
##  Max.   : 3.03250   Max.   : 3.273354   Max.   : 2.971878
```

The data consists of 500 rows and 12 columns. Let us conduct the exploratory factor analysis to learn about the factor structure of the data:

```r
analysis <- factanal(df2, factors = 4)
print(analysis)
```

```
##
## Call:
## factanal(x = df2, factors = 4)
##
## Uniquenesses:
##    y1    y2    y3    y4    y5    y6    y7    y8    y9   y10   y11   y12
## 0.588 0.346 0.594 0.581 0.424 0.543 0.462 0.470 0.498 0.520 0.376 0.559
##
## Loadings:
##     Factor1 Factor2 Factor3 Factor4
## y1                  0.637
## y2                  0.807
## y3                  0.632
## y4                          0.645
## y5                          0.757
## y6                          0.673
## y7   0.733
## y8   0.727
## y9   0.706
## y10          0.691
## y11          0.789
## y12          0.659
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      1.576   1.544   1.467   1.453
## Proportion Var   0.131   0.129   0.122   0.121
## Cumulative Var   0.131   0.260   0.382   0.503
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 25.36 on 24 degrees of freedom.
## The p-value is 0.386
```

Based on these results the data contains four factors. This is also supported by the p-value (chi square



statistic = 25.36, p = 0.386) which indicates that four factors are sufficient. A graph of the model is presented on top of the page.