

Exercises: R in Action

Ulrich Matter

Data: Student's GPA

The data used in the exercises are based on The data file is in sav-format (SPSS). We can read data from other stats packages into R with the **foreign** library. Thus, first, we load and install this R-package:

```
# install the package called "foreign" with the following command (if not yet installed)
# install.packages("foreign", repos = 'http://cran.us.r-project.org')
library(foreign)
```

Read the data into R as follows:

```
print(getwd())

## [1] "/Users/ueli/Dropbox/Teaching/Berkstats/Berkstats/notes"

sample <- read.spss("../data/sample_data.sav", to.data.frame = TRUE)
```

Have a look at the data set:

```
names(sample)

## [1] "SSATScore"          "ACTscore"
## [3] "HSGPA"              "SpringSemesterGPA"
## [5] "OverallGPA"         "CreditsatUniv"
## [7] "ClassPrepTime"      "CocurricularActTime"
## [9] "mult_classFB"       "mult_classTwitter"
## [11] "mult_classIM"       "mult_classEmail"
## [13] "mult_classSearch"   "mult_classTexting"
## [15] "sex"                "latino"
## [17] "race"               "OnOffCampusResidence"
## [19] "maxhighested"       "male"
## [21] "female"             "africanamerican"
## [23] "asianamerican"      "other"
## [25] "white"              "latinodv"
## [27] "lthighschool"       "highschool"
## [29] "somecollege"        "collegegrad"
## [31] "gradstudy"          "internetskills"
## [33] "facebookminutesselfreport"

View(sample)
```

“Clean” the data

```
sample <- sample[!is.na(sample$SSATScore),] # remove observations without SSATScore
sample <- sample[!is.na(sample$HSGPA),] # remove observations without HSGPA
sample <- sample[!is.na(sample$race),] # remove observations without race attribute
sample <- sample[!is.na(sample$sex),] # remove observations without gender attribute
```

Descriptives

Average time studying and average college GPA

```
mean(sample$ClassPrepTime)
```

```
## [1] 13.70533
```

```
mean(sample$OverallGPA)
```

```
## [1] 3.382163
```

Percentages of female, African-American?

```
mean(sample$female) * 100
```

```
## [1] 71.15987
```

```
mean(sample$africanamerican) * 100
```

```
## [1] 8.15047
```

Correlations

What is the relationship between SAT Score and College GPA?

```
cor(x = sample$SSATScore, y = sample$OverallGPA)
```

```
## [1] 0.323628
```

What is the relationship between High School GPA and College GPA?

```
cor(x = sample$HSGPA, y = sample$OverallGPA)
```

```
## [1] 0.3391248
```

Which one predicts more of the variance in College GPA?

Hypothesis test

Is there a difference in SAT scores between men and women?

```
anova(lm(SSATScore~factor(sex), data=sample)) # anova
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: SSATScore
```

```
##          Df    Sum Sq Mean Sq F value Pr(>F)
```

```
## factor(sex)  1      19316    19316   0.2678  0.6052
```

```
## Residuals  317  22864521    72128
```

```
summary(lm(SSATScore~factor(sex), data=sample)) # t-test of regression coefficient
```

```
##
```

```
## Call:
```

```
## lm(formula = SSATScore ~ factor(sex), data = sample)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -1803.93  -128.35    38.07   170.24   485.24
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          1831.93      28.00  65.426   <2e-16 ***
## factor(sex)Female    -17.18      33.19  -0.517    0.605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 268.6 on 317 degrees of freedom
## Multiple R-squared:  0.0008441, Adjusted R-squared:  -0.002308
## F-statistic: 0.2678 on 1 and 317 DF,  p-value: 0.6052
```

Is there a difference in College GPA between men and women?

```
anova(lm(OverallGPA~factor(sex), data=sample)) # anova
```

```
## Analysis of Variance Table
##
## Response: OverallGPA
##              Df Sum Sq Mean Sq F value    Pr(>F)
## factor(sex)   1  1.558  1.55788   8.3751 0.004068 **
## Residuals    317 58.966  0.18601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(OverallGPA~factor(sex), data=sample)) # t-test of regression coefficient
```

```
##
## Call:
## lm(formula = OverallGPA ~ factor(sex), data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63239 -0.25452  0.06335  0.32835  0.72761
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.27239     0.04497   72.776 < 2e-16 ***
## factor(sex)Female  0.15426     0.05330    2.894  0.00407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4313 on 317 degrees of freedom
## Multiple R-squared:  0.02574, Adjusted R-squared:  0.02267
## F-statistic: 8.375 on 1 and 317 DF,  p-value: 0.004068
```

Are there differences in College GPA among members of different racial groups?

```
anova(lm(OverallGPA~factor(race), data=sample)) # anova
```

```
## Analysis of Variance Table
##
## Response: OverallGPA
##              Df Sum Sq Mean Sq F value    Pr(>F)
## factor(race)   3  1.922  0.64070   3.4439 0.01708 *
## Residuals    315 58.602  0.18604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(OverallGPA~factor(race), data=sample)) # t-test of regression coefficient
```

```
##
## Call:
## lm(formula = OverallGPA ~ factor(race), data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68698 -0.25246  0.08316  0.29316  0.80192
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.34200    0.07291  45.839  <2e-16 ***
## factor(race)African American -0.18392    0.11167  -1.647    0.101
## factor(race)Asian      -0.01502    0.09819  -0.153    0.878
## factor(race)White       0.08484    0.07862   1.079    0.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4313 on 315 degrees of freedom
## Multiple R-squared:  0.03176,    Adjusted R-squared:  0.02254
## F-statistic: 3.444 on 3 and 315 DF,  p-value: 0.01708
```

Multiple regression

Controlling for pre-existing ability, are there differences in College GPA among members of different racial groups?

```
anova(lm(OverallGPA~factor(race) + HSGPA, data=sample)) # anova
```

```
## Analysis of Variance Table
##
## Response: OverallGPA
##              Df Sum Sq Mean Sq F value    Pr(>F)
## factor(race)   3  1.922  0.6407   3.8551 0.00986 **
## HSGPA           1  6.417  6.4168 38.6098 1.643e-09 ***
## Residuals     314 52.185  0.1662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(OverallGPA~factor(race) + HSGPA, data=sample)) # t-test of regression coefficient
```

```
##
## Call:
## lm(formula = OverallGPA ~ factor(race) + HSGPA, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64261 -0.26021  0.06861  0.28097  0.95196
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.54699    0.29699   5.209 3.44e-07 ***
## factor(race)African American -0.16705    0.10558  -1.582    0.115
```

```
## factor(race)Asian          -0.08348    0.09346  -0.893    0.372
## factor(race)White          0.04094    0.07464   0.548    0.584
## HSGPA                      0.44368    0.07140   6.214 1.64e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4077 on 314 degrees of freedom
## Multiple R-squared:  0.1378, Adjusted R-squared:  0.1268
## F-statistic: 12.54 on 4 and 314 DF,  p-value: 1.766e-09
```

Does time spent on Facebook predict Overall College GPA when controlling for sex, race, and prior academic ability?

```
anova(lm(OverallGPA~factor(race) + factor(sex) + HSGPA + facebookminutesselfreport, data=sample)) # ano
```

```
## Analysis of Variance Table
##
## Response: OverallGPA
##              Df Sum Sq Mean Sq F value    Pr(>F)
## factor(race)    3  1.922   0.6407   3.9711 0.008445 **
## factor(sex)      1  1.768   1.7676  10.9555 0.001043 **
## HSGPA            1  5.962   5.9621  36.9537 3.527e-09 ***
## facebookminutesselfreport 1  0.534   0.5342   3.3112 0.069766 .
## Residuals      312 50.338   0.1613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(OverallGPA~factor(race) + factor(sex) + HSGPA + facebookminutesselfreport, data=sample)) # t
```

```
##
## Call:
## lm(formula = OverallGPA ~ factor(race) + factor(sex) + HSGPA +
##     facebookminutesselfreport, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56117 -0.23749  0.07998  0.27615  0.85094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.5573570   0.2940925   5.295 2.24e-07 ***
## factor(race)African American -0.1799403   0.1042565  -1.726  0.08535 .
## factor(race)Asian          -0.0736564   0.0923476  -0.798  0.42571
## factor(race)White           0.0471087   0.0736462   0.640  0.52286
## factor(sex)Female           0.1533837   0.0502575   3.052  0.00247 **
## HSGPA                0.4259558   0.0705633   6.037 4.47e-09 ***
## facebookminutesselfreport  -0.0003181   0.0001748  -1.820  0.06977 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4017 on 312 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.1523
## F-statistic: 10.52 on 6 and 312 DF,  p-value: 1.225e-10
```