Applied Data Analysis, Übung 4

Henry Haustein

Vorbereitung

```r
install.packages("dplyr")
library(dplyr)
install.packages("readxl")
library(readxl)
data = read_excel("data.xlsx", na = "NA")

data$gender = factor(data$gender)
levels(data$gender) = c("male", "female", "diverse")
data$employment = factor(data$employment)
levels(data$employment) = c("student", "employed", "unemployed")
data$education = factor(data$education)
levels(data$education) = c("no degree", "secondary", "intermediate
    ", "high school", "academic")
data$play_frequency = factor(data$play_frequency)
levels(data$play_frequency) = c("never", "every few months", "
    every few weeks", "1-2 days a week", "3-5 days a week", "daily
    ")
data$treatment = factor(data$treatment)
levels(data$treatment) = c("control", "lootbox in task reward", "
    lootbox picture", "badge")
data$age = sapply(data$age, function(year) {2016-year})
data$rt6 = as.numeric(data$rt6)
data$rt7 = as.numeric(data$rt7)
data$rt8 = as.numeric(data$rt8)
data$rt9 = as.numeric(data$rt9)
data$rt10 = as.numeric(data$rt10)
data$rt11 = as.numeric(data$rt11)
data$rt12 = as.numeric(data$rt12)
data$rt13 = as.numeric(data$rt13)
data$rt14 = as.numeric(data$rt14)
subControl = subset(data, treatment == "control")
```

Task 1

```r
install.packages("ggplot2")
library(ggplot2)
ggplot(data, aes(x = age, y = `total time`))
```

```
+ geom_point()
+ geom_smooth(method = "lm")
lm_tt_age = lm(`total time` ~ age, data = data)
summary(lm_tt_age)
ggplot(subControl, aes(x = age, y = `total time`))
+ geom_point()
+ geom_smooth(method = "lm")
lm_tt_age_control = lm(`total time` ~ age, data = subControl)
summary(lm_tt_age_control)

ggplot(data, aes(x = tasks_completed, y = `total time`))
+ geom_point()
+ geom_smooth(method = "lm")
lm_tt_tc = lm(`total time` ~ tasks_completed, data = data)
summary(lm_tt_tc)
ggplot(subControl, aes(x = tasks_completed, y = `total time`))
+ geom_point()
+ geom_smooth(method = "lm")
lm_tt_tc_control = lm(`total time` ~ tasks_completed, data =
    subControl)
summary(lm_tt_tc_control)

ggplot(data, aes(x = rt0, y = `total time`))
+ geom_point()
+ geom_smooth(method = "lm")
lm_tt_rt0 = lm(`total time` ~ rt0, data = data)
summary(lm_tt_rt0)
ggplot(subControl, aes(x = rt0, y = `total time`))
+ geom_point()
+ geom_smooth(method = "lm")
lm_tt_rt0_control = lm(`total time` ~ rt0, data = subControl)
summary(lm_tt_rt0_control)
```

Das führt zu folgenden Modellen:

	Full Model	Control Group
(Intercept)	230665.09***	298612.42
	(53287.10)	(156127.98)
age	3635.06	−1457.79
	(2032.53)	(5666.45)
R^2	0.01	0.00
Adj. R^2	0.01	−0.01
Num. obs.	437	99

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

	Full Model	Control Group
(Intercept)	170842.92***	280409.31*
	(44899.68)	(128200.11)
tasks_completed	14859.12***	−2518.17
	(4095.99)	(15008.84)
R^2	0.03	0.00
Adj. R^2	0.03	−0.01
Num. obs.	437	99

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

	Full Model	Control Group
(Intercept)	157695.17***	119595.89***
	(5821.01)	(5304.52)
rt0	1.01***	0.99***
	(0.01)	(0.01)
R^2	0.91	0.99
Adj. R^2	0.91	0.99
Num. obs.	437	99

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Man sieht, dass offensichtlich nur die Variable `rt0` die Variable `total time` gut erklären kann.

Task 2

```
lm_tt_age_rt0 = lm(‘total time‘ ~ age + rt0, data = data)
summary(lm_tt_age_rt0)

lm_tt_age_rt0_interaction = lm(‘total time‘ ~ age + rt0 + age:rt0,
     data = data)
summary(lm_tt_age_rt0_interaction)

lm_tt_treatment_rt0 = lm(‘total time‘ ~ treatment + treatment:rt0
    + 0, data = data)
summary(lm_tt_treatment_rt0)
```

Das führt zu folgenden Modellen:

	without interaction terms	with interaction terms
(Intercept)	107465.81***	115139.46***
	(15600.65)	(20861.24)
age	2048.80***	1700.27*
	(591.55)	(863.35)
rt0	1.01***	0.98***
	(0.01)	(0.06)
age:rt0		0.00
		(0.00)
R^2	0.92	0.92
Adj. R^2	0.92	0.92
Num. obs.	437	437

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

	Model 1
treatmentcontrol	119595.89***
	(10859.57)
treatmentlootbox in task reward	224714.96***
	(12861.18)
treatmentlootbox picture	135344.63***
	(18221.65)
treatmentbadge	153816.42***
	(10165.92)
treatmentcontrol:rt0	0.99***
	(0.02)
treatmentlootbox in task reward:rt0	1.03***
	(0.04)
treatmentlootbox picture:rt0	1.06***
	(0.13)
treatmentbadge:rt0	1.01***
	(0.02)
R^2	0.96
Adj. R^2	0.96
Num. obs.	437

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Sobald wir Interaktionsterme erlauben, ist das Alter der Probanden gar nicht mehr so wichtig. Die Behandlung scheint aber einen großen Einfluss auf die `total time` zu haben.