Practical Exercise 3 | Statistics for CSAI II

Burcu Ibicioglu_ u986202

The goals of this exercise are to (a) to use R to run multiple linear regression models, b) check the assumptions of the model and c) report your results.

Part A

For this part of Practical Exercise #3, Tasks indicate things that you need to complete in R/R Studio.

Task 1. Load data from the "Goggles Regression" data set (GogglesRegression.dat).

```
data <- read.table('Goggles.dat', header = TRUE)
```

Task 2. Inspect the data by looking at the first few entries and the last few entries in the dataset. This data is meant to examine the "Beer Goggles" prediction that someone will rate others as more attractive after consuming alcohol. Alcohol indicates whether the person had no alcohol (0) or drank alcohol (1). Attractiveness represents a rating of attractiveness. Gender is male (0) and female (1). Interaction is the product of gender and alcohol.

```
head(data)
```

```
##   gender alcohol interaction attractiveness
## 1      1       0           0             65
## 2      1       0           0             70
## 3      1       0           0             60
## 4      1       0           0             60
## 5      1       0           0             60
## 6      1       0           0             55
```

```
tail(data)
```

```
##    gender alcohol interaction attractiveness
## 27      0       1           0             30
## 28      0       1           0             55
## 29      0       1           0             35
## 30      0       1           0             20
## 31      0       1           0             45
## 32      0       1           0             40
```

a. Generate descriptive statistics by alcohol groups. Evaluate these descriptives and print them here.

```
install.packages("psych")
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```r
describeBy(data$alcohol)
```

```
## Warning in describeBy(data$alcohol): no grouping variable requested

##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 32  0.5 0.51    0.5     0.5 0.74   0   1     1    0    -2.06 0.09
```

b. Generate a correlation matrix for all variables in the data set and print it here.

```r
cm <- cor(data)
cm
```

```
##                   gender    alcohol interaction attractiveness
## gender         1.0000000  0.0000000  0.57735027     0.27357145
## alcohol        0.0000000  1.0000000  0.57735027    -0.60185718
## interaction    0.5773503  0.5773503  1.00000000     0.09476793
## attractiveness 0.2735714 -0.6018572  0.09476793     1.00000000
```

c. State a null hypothesis and an alternative hypothesis about the relationship between alcohol and attractiveness ratings, as well as gender and attractiveness ratings.

Null: There is no relationship between alcohol and attractiveness ratings Alternative: There is a relationship between alcohol and attractiveness ratings

Null: There is no relationship between gender and attractiveness ratings Alternative: There is a relationship between gender and attractiveness ratings

Task 3. Run a multiple regression model that includes alcohol and gender as predictors of attractiveness. Generate 95% confidence of the b estimates and also generate the standardized beta estimates.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
model <- lm(attractiveness ~ alcohol + gender, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = attractiveness ~ alcohol + gender, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6563  -7.6562  -0.4687   6.2500  20.1563
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.844      3.446  17.368  < 2e-16 ***
## alcohol      -17.187      3.979  -4.320 0.000167 ***
## gender         7.812      3.979   1.964 0.059234 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.25 on 29 degrees of freedom
## Multiple R-squared:  0.4371, Adjusted R-squared:  0.3983
## F-statistic: 11.26 on 2 and 29 DF,  p-value: 0.0002407
```

```r
ci<- confint(model, level = 0.95)
ci
```

```
##                   2.5 %    97.5 %
## (Intercept)  52.796513 66.890987
## alcohol     -25.324948 -9.050052
## gender       -0.324948 15.949948
```

```r
beta <- coef(model)
beta
```

```
## (Intercept)     alcohol      gender
##    59.84375   -17.18750     7.81250
```

a. Report the results here in APA format. Be sure to include the adjusted R2 value, the b estimates, the p-values, and the 95% confidence intervals. What can you conclude from your results?

The regression analysis that was conducted to test the relationship between alcohol consumption and gender as predictors of attractiveness ratings shows that the model was significant, since the p-value is 0.0002407,which means p<0.001. The adjusted R-squared is 0.3983.

b. Write your results here in equation form.

#Hint: an easy guide to mathematical expression in LaTeX is here: https://www.overleaf.com/learn/latex/mathematical_expressions

```r
"Attractiveness = 59.844-17.18750(Alcohol) + 7.81250(Gender)"
```

```
## [1] "Attractiveness = 59.844-17.18750(Alcohol) + 7.81250(Gender)"
```

Task 4. Run a multiple regression model that includes alcohol and gender as well as the interaction term of both variables as predictors of attractiveness. Generate 95% confidence of the b estimates and also generate the standardized beta estimates.

```r
m <- lm(attractiveness ~ alcohol * gender, data = data)

confint(m, level = 0.95)
```

```
##                     2.5 %     97.5 %
## (Intercept)      60.61708  73.132925
## alcohol         -40.10004 -22.399958
## gender          -15.10004   2.600042
## alcohol:gender   15.60915  40.640850
```

```r
modelst <- lm(scale(attractiveness) ~ scale(alcohol) * scale(gender), data = data)

confint(modelst, level = 0.95)
```

```
##                                     2.5 %     97.5 %
## (Intercept)                   -0.21568351  0.2156835
## scale(alcohol)                -0.82099184 -0.3827225
## scale(gender)                  0.05443678  0.4927061
## scale(alcohol):scale(gender)   0.27766692  0.7229490
```

a. Report the results here in APA format. Be sure to include the adjusted R2 value, the b estimates, the p-values, and the 95% confidence intervals. What can you conclude from your results?

A multiple regression analysis was used to spot the impact of alcohol consumption, gender, and their interaction with the attractiveness ratings. The model was significant, since R-squared = 0.4 p< 0.001

Task 5. Find out whether the second model which includes the interaction term is a better fit of the data than the model that does not include the interaction term. Use anova() and extractAIC().

```r
ni <- lm(attractiveness ~ alcohol + gender, data = data)
i <- lm(attractiveness ~ alcohol * gender, data = data)

anova(ni, i)
```

```
## Analysis of Variance Table
##
## Model 1: attractiveness ~ alcohol + gender
## Model 2: attractiveness ~ alcohol * gender
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 3672.7
## 2     28 2090.6  1      1582 21.188 8.204e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ani <- extractAIC(ni)
ai <- extractAIC(i)
```

a. Report the change in the percentage of variability accounted for by the better fitting model.

The interaction in Model 2 has more variability in attractiveness compared to Model 1. The difference is statistically significant, since F(1,28)=21.19 and p<.001. So, the interaction between alcohol and gender improves the model's fit.

Task 6. To better understand the significant interaction, we need to generate simple slopes for the model. What are the simple intercepts and simple slope for the model? Write them out in equation form and make a graph that shows how the slope changes at the various levels of gender.

```r
"for females: attractiveness = 60.62-40.10(alcohol)"
```

```
## [1] "for females: attractiveness = 60.62-40.10(alcohol)"
```

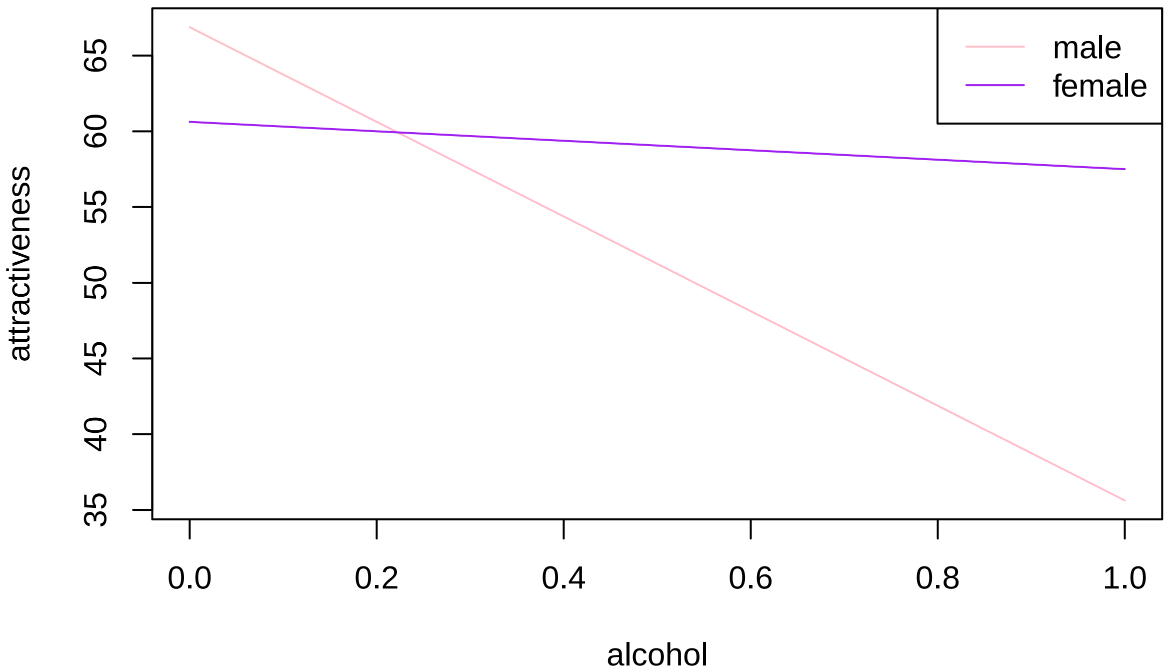
```r
"for males: attractiveness = 45.52-24.49(alcohol)"
```

```
## [1] "for males: attractiveness = 45.52-24.49(alcohol)"
```

```r
model <- lm(attractiveness ~ alcohol * gender, data = data)
as<- seq(0, max(data$alcohol), length.out = 100)

g0<- predict(model, newdata = data.frame(alcohol = as, gender = 0))
g1 <- predict(model, newdata = data.frame(alcohol = as, gender = 1))

plot(as, g0, type = "l", col = "pink", xlab = "alcohol", ylab = "attractiveness")
lines(as, g1, col = "purple")
legend("topright", legend = c("male", "female"), col = c("pink", "purple"), lty = 1)
```



Task 7. Let's make sure this better fitting model doesn't violate any assumptions. Use gvlma() to do a global evaluation of the model (here we are only using gvlma because of the categorical predictors).

```
installed.packages('gvlma')
```

```
##      Package LibPath Version Priority Depends Imports LinkingTo Suggests
##      Enhances License License_is_FOSS License_restricts_use OS_type Archs
##      MD5sum NeedsCompilation Built
```

```
library(gvlma)

model <- lm(attractiveness ~ alcohol * gender, data = data)

r <- gvlma(model)

r
```

```
##
## Call:
## lm(formula = attractiveness ~ alcohol * gender, data = data)
##
## Coefficients:
##    (Intercept)         alcohol          gender  alcohol:gender
##          66.87          -31.25           -6.25           28.13
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model)
##
##                         Value p-value                Decision
```

```
## Global Stat        2.981e+00 0.56109 Assumptions acceptable.
## Skewness           1.417e-01 0.70665 Assumptions acceptable.
## Kurtosis           1.501e-02 0.90249 Assumptions acceptable.
## Link Function      1.681e-15 1.00000 Assumptions acceptable.
## Heteroscedasticity 2.824e+00 0.09287 Assumptions acceptable.
```