1.1

The UOI inventory is a test developed to measure students' success in speaking skills and is composed by twelve items on a 21-point rating scale. In a recent research, it has been administered to a sample of $n = 125$ high-school students along with a demographic sheet collecting information about gender, age, and gross annual income in Euro, which has been codified using three categories (i.e., a. $[0 - 15000]$; b. $]15000 - 30000]$; c. $]30000 - \text{higher}]$). The goal is to study whether the UOI total score - which has been computed as average of the twelve items - can be modeled as a function of the demographic variables.

1. Identify the number of statistical units, the number of variables for the analysis by distinguishing between the outcome variable (Y) and the predictors $(X_1, \ldots, X_J)$.
2. Identify the support of the outcome variable Y.
3. Write the linear function connecting Y to $X_1, \ldots, X_J$ and define the corresponding most appropriate statistal model (e.g., see slides B:17-18).

SOLUTION

1. The number of statistical units is $n = 125$, the number of variables available is $J + 1 = 4$, with UOI total score being the outcome Y. The remaining variables gender (X_1), age (X_2), and income (X_3) will be used as predictors.
2. The outcome variable is computed by applying the mean function over twelve variables (items) bounded in the interval $\{1, \ldots, 21\}$. Then, the support is $\sup(Y) = [1, 21]$, which is a bounded subset of real numbers.
3. The linear function is

$$\texttt{UOI} = \beta_0 + \texttt{gender}\beta_1 + \texttt{age}\beta_2 + \texttt{income}\beta_3 + \epsilon$$

In this case, the Normal linear model can be considered as a good candidate to this purpose:

$$\begin{aligned} y_i &\sim \mathcal{N}(y; \mu_i, \sigma^2) \quad i = 1, \ldots, 125 \\ \mu_i &= \beta_0 + \mathbf{x}_i\boldsymbol{\beta} \end{aligned}$$

where $\mathbf{x}_i$ is a $1 \times J$ vector of predictors for the i-th observation, whereas $\boldsymbol{\beta}$ is a vector of appropriate order. However, since we are dealing with bounded outcome variables, the adequacy of this model should carefully be checked once the parameters will have been estimated.

1.2

Consider an experiment to study memory in a clinical sample of $n = 78$ pre-school children where the following variables have been collected: (i) number of fails in recognizing the stimulus, (ii) age in months, (iii) neurological impairment as measured by standardized values via the RYY test, (iv) number of weeks from the latest neurological episode. The goal is to study whether experiment failures can be predicted by the other variables.

1. Identify the number of statistical units, the outcome variable Y, and the predictors $X_1, \ldots, X_J$.
2. Identify the support of the outcome variable and the most appropriate statistical (linear) model for the current analysis.

3. Identify the number of parameters the model takes along with the parameter space.

SOLUTION

1. The number of statistical units is $n = 78$, the outcome is the number of errors (Y) in the recognition task, the predictors are represented by age (X_1), neurological impairment (X_2), and number of weeks from the latest episode (X_3).

2. The outcome variable consists of counts, therefore $\sup(Y) = \mathbb{N}_0$. The Poisson linear model can be considered an adequate model to analyse these data:

$$y_i \sim \mathcal{P}oi(y; \lambda_i) \quad i = 1, \ldots, 78$$
$$\lambda_i = \exp(\beta_0 + \texttt{age}\beta_1 + \texttt{neuro}\beta_2 + \texttt{weeks}\beta_3)$$

   However, as the mean $\mathbb{E}[Y_i]$ and the variance $\mathbb{V}\text{ar}[Y_i]$ are the same in this model, attention should be paid to excess of zeros or overdispersion in the data (the Poisson model could not lack in flexibility).

3. The array of parameter is $\boldsymbol{\theta} = \{\beta_0, \beta_1, \beta_2, \beta_3\} \in \mathbb{R}^4$, with length $p = 4$ (i.e., number of parameters).

1.3

In a context-recall task, $n = 78$ participants have been equally assigned to a control group and an experimental group. During the experiment, reaction times (in logarithmic scale) and accuracies have been measured. The researcher wants to assess whether reaction times vary as a function of the manipulation task.

1. Identify the number of statistical units, the outcome variable Y, and the predictors $X_1, \ldots, X_J$.

2. Identify the support of the outcome variable and the most appropriate statistical (linear) model for the current analysis.

3. Identify the number of parameters the model takes along with the parameter space.

4. Provide an interpretation for the model parameters.

SOLUTION

1. The number of statistical units is $n = 78$, the outcome is the reaction time in log scale (Y), the predictor is the experimental manipulation codified as group assignment ($X \in \{1, 2\}$).

2. The outcome variable consists of log-times, therefore $\sup(Y) = \mathbb{R}$. The Normal linear model can be considered to analyse these data:

$$y_i \sim \mathcal{N}(y; \mu_i, \sigma^2) \quad i = 1, \ldots, 78$$
$$\mu_i = \beta_0 + \texttt{group}_i\beta_1$$

3. The parameters are $\boldsymbol{\theta} = \{\beta_0, \beta_1, \sigma^2\} \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$. The number of parameters is $p = 3$.

4. As the predictor is categorical with two levels, the parameters $\{\beta_0, \beta_1\}$ codify the means for both groups as follows:

   β_0 : mean of $\texttt{group} = 1$
   β_1 : increment of the mean for $\texttt{group} = 2$ ($\texttt{group} = 1$ is the baseline)
   $\beta_0 + \beta_1$: mean of $\texttt{group} = 2$

1.4

Consider the dataset `teengamb` from the R library `faraway`. The dataset concerns a study of teenage gambling in Britain. First, transform the variable `gamble` as follows: `log.gamble = log(gamble+1)`. Second, make a numerical and graphical summary of the data in order to evaluate how `log.gamble` varies as a function of the other ones. Third, define and fit a Normal linear model to evaluate whether `log.gamble` is predicted by `sex`. The fit can be implemented by using either the ML solutions (see slides B:13,25,26) or the `lm()` function. Finally, provide comments for the estimated parameters and the overall fit of the model. Optionally: Plot the fitted model. For further information about the data, type `?faraway::teengamb` on the R console.

SOLUTION

```
## Load the data and check the data structure
datax = faraway::teengamb
str(datax)
datax$sex = as.factor(datax$sex)  #sex should be transformed as a categorical variable
datax$log.gamble = log(datax$gamble + 1)  #variable transformation as required

## Descriptive analyses on the data
psych::describe(x = datax[, -5])  #the original variable should be removed from the analysis
psych::describeBy(x = datax$log.gamble, group = datax$sex)
psych::pairs.panels(x = datax[, c(2:4, 6)])  #it works for numeric variables only
boxplot(datax$log.gamble[datax$sex == 0], datax$log.gamble[datax$sex == 1], frame = FALSE,
    names = c("0", "1"))
```

The Normal linear model is defined as follows:

$$\begin{aligned}&\texttt{log.gamble}_i \sim \mathcal{N}(y; \mu_i, \sigma^2) \quad i = 1, \ldots, 47\\ &\mu_i = \beta_0 + \texttt{sex}_i \beta_1\end{aligned}$$

whereas parameter estimation and fit evaluation can be performed using R.

```
## Fit using the ML solutions directly
X = model.matrix(~datax$sex)  #it creates the X matrix of predictors
y = datax$log.gamble  #outcome
n = NROW(datax)  #number of observations
J = 1  #number of predictor
b_est = solve(t(X) %*% X) %*% t(X) %*% y  #beta
sigma_est = 1/n * t(y - X %*% b_est) %*% (y - X %*% b_est) * (n/(n - J - 1))  #sigma^2
se_b_est = sqrt(diag(solve(t(X) %*% X)) * sigma_est)
y_hat = X %*% b_est
one = matrix(data = 1, nrow = n, ncol = 1)
r_squared = 1 - ((t(y - y_hat) %*% (y - y_hat))/(t(y - one * mean(y)) %*% (y - one *
    mean(y))))  #unadjusted index

print(b_est)
print(sigma_est)
print(se_b_est)
print(r_squared)

## Fit using the lm() function
lm_def = as.formula(log.gamble ~ sex)  #model definition
lm_fit = lm(formula = lm_def, data = datax)  #model fit
out = summary(lm_fit)
print(out)

# extract the estimated quantities
b_est = out$coefficients[, 1]
se_b_est = out$coefficients[, 2]
sigma_est = out$sigma^2
r_squared = out$r.squared
```

```
print(b_est)
print(sigma_est)
print(se_b_est)
print(r_squared)
```

The results suggest that `log.gamble` decreased as a function of `sex`: In particular participants in the group `sex=0` showed a higher `log.scale` as opposed to participants in the group `sex=1`. The estimated mean of the group `sex=0` is equal to $\beta_0 = 2.555$ whereas the estimated mean of the remaining group is $\beta_0 + \beta_1 = 2.555 - 1.444 = 1.111$.

Finally, the optional plot can be get using the following R commands:

```
## Fit the model
lm_def = as.formula(log.gamble ~ sex)  #model definition
lm_fit = lm(formula = lm_def, data = datax)  #model fit
out = summary(lm_fit)
b_est = out$coefficients[, 1]

## Plot the observed points
plot(x = rep(1, length(datax$log.gamble[datax$sex == 0])), datax$log.gamble[datax$sex ==
    0], xlim = c(0.5, 2), bty = "n", xlab = "sex", ylab = "log.gamble", pch = 20)
points(x = rep(1.5, length(datax$log.gamble[datax$sex == 1])), datax$log.gamble[datax$sex ==
    1], bty = "n", pch = 20)

## Plot the estimated means
points(x = 1, y = b_est[1], pch = 17, col = 2, cex = 2)  #sex=0
points(x = 1.5, y = b_est[1] + b_est[2], pch = 17, col = 2, cex = 2)  #sex=1
abline(a = b_est[1], b = b_est[2], col = "darkgray", lty = 2, lwd = 2)
```

1.5

The dataset `uswages` from the library `faraway` contains data about weekly wages for US male workers from the Current Population Survey (cohort year: 1988). The goal is to define and fit a linear model to predict weekly wages (in log scale) as a function of years of education and years of experience. For further information about the data, type `?faraway::uswages` on the R console.

1. Identify the number of statistical units, the outcome variable Y, and the predictors $X_1, \dots, X_J$.

   The number of statistical units is $n = 2000$, the outcome is the weekly wage (Y), the predictors are the years of education and the year of work experience (note that the dataset contains other variables that might be used as further predictors).

2. Identify the support of the outcome variable and the most appropriate statistical linear model for the current analysis.

   The outcome variable consists of weekly wages in log scale, therefore $\sup(Y) = \mathbb{R}$. The Normal linear model can be considered to analyse these data:

   $$y_i \sim \mathcal{N}(y; \mu_i, \sigma^2) \quad i = 1, \dots, 78$$
   $$\mu_i = \beta_0 + \texttt{educ}_i\beta_1 + \texttt{exper}_i\beta_2$$

3. Identify the number of parameters the model takes along with the parameter space.

   The parameters are $\boldsymbol{\theta} = \{\beta_0, \beta_1, \beta_2, \sigma^2\} \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$. The number of parameters is $p = 4$.

4. Draw a random sample of $n = 120$ from the dataset `uswages`. To do so, use the function `random_subsample(X,n = 120,seedx = 122)` with `X` being the current dataset. Note that the function can be loaded by typing `source("utilities.R")`. The file `utilities.R` is available on the folder "Datasets & Utilities". From now on, use the subset of data for the further analyses.

```
datax = faraway::uswages
source("../labs/utilities.R")  #use your own directory and path
datax = random_subsample(X = datax, n = 120, seedx = 122)
```

5. Define a new variable as follows: `logwage = log(wage)`.

```
datax$logwage = log(datax$wage)
```

6. Make a graphical summary of the relationships among `logwage`, `edu`, and `exper`.

```
par(mfrow = c(1, 2))
plot(datax$educ, datax$logwage, bty = "n", xlab = "educ", ylab = "log(wage)", pch = 20)
plot(datax$exper, datax$logwage, bty = "n", xlab = "exper", ylab = "log(wage)", pch = 20)
```

7. Define and fit an appropriate linear model in order to predict `logwage` as a function of `edu` and `exper`.

```
out = lm(formula = logwage ~ educ + exper, data = datax)
```

8. Evaluate the overall fit of the model and give an interpretation to the regression coefficients. Plot the estimated model.

```
summary(out)
plot(effects::allEffects(out))
```

The fit of the current model is not satisfactory at all. Overall, it explains about the 20% of the observed variability of Y. The remaining amount of variability is not currently explained by the predictors. This suggests that further variables might be added to the model in order to explain the variability of weekly wages.
With regards to the current regression coefficients we can state as follows: A unit variation of `educ` increases `logwage` by 0.119 whereas a unit variation of `exper` increases `logwage` by 0.021. However, as the outcome variable has been log transformed, it is natural to interpret the exponentiated regression coefficients. These values correspond to changes in the ratio of the expected geometric means of the original outcome variable (e.g., `wage`). In particular, each unit of `educ` increases the untransformed `wage` by $\exp(\beta_1 = 0.1196) = 1.127$ or we expect to see a $(\exp(0.1196) - 1)100) = 12.7\%$ of variation in weekly wages. Similarly, each unit increase of `exper` increases the untransformed `wage` by $\exp(\beta_2 = 0.021) = 1.021$ or we expect to see a $(\exp(0.021) - 1)100) = 2.17\%$ of variation in weekly wages.

9. Compute the 90% confidence interval associated with the predictors. Using just these intervals, what can we conclude about the coefficients at the population level?

```
confint(out, level = 0.9)
```

The CIs for the regression parameters do not contain the null value zero. We can conclude that both coefficients significantly differ by zero ($\alpha = 0.10$). We also notice that CI(`educ`) is slightly larger than CI(`exper`).

10. Consider the estimated coefficients. In order to get a higher weekly wage, would you suggest to your children to get a higher education level or, by contrast, to get more work experience? Please, justify why. We will suggest to get a higher education level. This is due to the fact that, based on the sample data being analysed, the variable `educ` produces a larger increase of the outcome (about 12% per unit).

11. Use the incremental-F test procedure in order to check whether the current model with two predictors is really the best model given the available data (see `lab2.R`, section A.4).

```r
# Note: The variable with higher F-value has to be included.
mod0 = lm(formula = logwage ~ 1, data = datax)
add1(mod0, scope = out, test = "F")

mod0 = lm(formula = logwage ~ educ, data = datax)
add1(mod0, scope = out, test = "F")
```

1.6

Consider the dataset `dataex1.csv` on the folder "Datasets & Utilities". It contains five continuous variables `X1`,..,`X5` as well as a continuous response variable `y`. The goal is to define and fit a Normal linear model by selecting the best subset of predictors explaining the outcome. Note: files with extension `.csv` can be loaded into R by using the function `read.csv(file = "namepath/dataex1.csv",header = TRUE)` with `namepath` being the current path locating the file `dataex1.csv` on your local machine.

1. Load the data.

   ```r
   datax = read.csv(file = "dataex1.csv", header = TRUE)
   ```

2. Make a numerical and graphical summary of the relationships among `y` and the predictors.

   ```r
   psych::describe(x = datax)
   psych::pairs.panels(datax)
   ```

3. Use a statistical procedure to select the best subset of predictors for the outcome `y`.

   ```r
   mod_full = lm(formula = y ~ ., data = datax)  #complete model
   mod0 = lm(formula = y ~ 1, data = datax)  #null model

   add1(mod0, scope = mod_full, test = "F")
   # X4 has to be retained for the next analysis

   mod0 = lm(formula = y ~ X4, data = datax)
   add1(mod0, scope = mod_full, test = "F")
   # X2 has to be retained for the next analysis

   mod0 = lm(formula = y ~ X4 + X2, data = datax)
   add1(mod0, scope = mod_full, test = "F")
   # X3 has to be retained for the next analysis

   mod0 = lm(formula = y ~ X4 + X2 + X3, data = datax)
   add1(mod0, scope = mod_full, test = "F")
   # The remaining variables do not contribute to significantly change the fit of
   # the model.  The procedure stops here.

   mod_final = lm(formula = y ~ X4 + X2 + X3, data = datax)
   ```

   The F-test based procedure has selected three variables `X2`, `X3`, `X4` out of five potential predictors.

4. Describe the results of the final model. Use the partial regression plots to simplify the interpretation of the parameters.

   ```r
   summary(mod_final)
   car::avPlots(mod_final, id = FALSE)
   ```

The final model explains about 0.90% of the variance of `y`. Two of the current predictors are positively related, namely `X2` ($\beta_{\texttt{X2}} = 2.624$, $\sigma_{\beta_{\texttt{X2}}} = 0.1518$) and `X3` ($\beta_{\texttt{X3}} = 0.6586$, $\sigma_{\beta_{\texttt{X3}}} = 0.1616$), whereas one of them is negatively related to the outcome, i.e. `X4` ($\beta_{\texttt{X4}} = -3.864$, $\sigma_{\beta_{\texttt{X4}}} = 1515$). The t-statistics associated to the regression coefficients are significant ($\alpha = 0.05$), with $t_{\beta_{\texttt{X4}}}$ showing the largest value.

5. Consider the F statistic of the omnibus test. Plot the probability distribution of the statistic test F with degrees of freedom equal to those reported in the summary of the fitted model. Where is the observed F-statistic (i.e., $W = 289.10$) located in this plot?

```
curve(expr = df(x, df1 = 3, df2 = 100 - 5 - 1), 0, 10, bty = "n")
```

Considering the plot of the F-distribution, the observed statistic is located on the tail of the distribution. Its observed value is extreme if compared to a F-distribution with these degree of freedom. This is why the probability associated to the statistic is closed to zero.

6. Consider the parameter $\beta_{\texttt{X2}}$. Test the hypothesis that $H_0 : \beta_{\texttt{X2}} = 2.30$ against $H_0 : \beta_{\texttt{X2}} \neq 2.30$ (use $\alpha = 0.01$).

```
n = NROW(datax)
beta_est = summary(mod_final)$coefficients[3, 1]
sd_beta = summary(mod_final)$coefficients[3, 2]
tb = (beta_est - 2.3)/sd_beta
p_tb = 2 * min(pt(q = -tb, df = n - 3 - 1), pt(q = tb, df = n - 3 - 1))  #p-value
print(p_tb)
```

The probability $\mathbb{P}(T \geq \texttt{tb}) = 0.0349$ (p-value) is higher then the fixed level α. Then, there is no evidence against H_0: The estimated $\beta_{\texttt{X2}}$ does not significantly differ from $\beta^0 = 2.30$ at the population level.

1.7

Consider the dataset `datax2.csv` on the folder "Datasets & Utilities". It contains four continuous variables `X1`,..,`X4` as well as a continuous response variable `y`. The goal is to define and fit a Normal linear model by selecting the best subset of predictors explaining the outcome. Note: The file can be loaded using the function `read.csv(file = "namepath/datax2.csv",header = TRUE)` with `namepath` being the current path locating the file `datax2.csv` on your local machine.

1. Load the data.

```
datax = read.csv(file = "datax2.csv", header = TRUE)
source("../labs/utilities.R")
```

2. Make a numerical and graphical summary of the relationships among `y` and the predictors. Indicate the support of the outcome variable `y` and provide comments about the graphical results.

```
psych::describe(x = datax)
exploratory_plots(y = datax[, 1], X = datax[, 2:5], plot_type = "loess")
```

The support of the outcome variable is the continuous interval $[-14.87, 22.64]$. The outcome seems to be positively related to `X1` and negatively associated with `X2`. No linear relationships between the outcome and the remaining predictors `X2` and `X4` can be visually detected.

3. Select the best subset of predictors for the outcome variable by means of an appropriate procedure.

```
leaps_r2(y = datax[, 1], X = datax[, 2:5])
```

All the available predictors are continuous so that the function `leaps_r2()` can be used. It implements a procedure based on the maximization of the adjusted R2 index. The final model is that one, considering a set of potential models, maximizing the adjusted R2 index. Another possibility would be to use the incremental F-test by means of the function `add1()`, as follows. Note that, in this case, the results might be different.

```
full_model = lm(formula = y ~ ., data = datax)

current_model = lm(formula = y ~ 1, data = datax)
add1(scope = full_model, object = current_model, test = "F")

current_model = lm(formula = y ~ X1, data = datax)
add1(scope = full_model, object = current_model, test = "F")

current_model = lm(formula = y ~ X1 + X3, data = datax)
add1(scope = full_model, object = current_model, test = "F")

current_model = lm(formula = y ~ X1 + X3 + X4, data = datax)
add1(scope = full_model, object = current_model, test = "F")
```

4. Define and fit a Normal linear model considering the best subset of predictors.

   The Normal linear model is as follows:

$$y_i \sim \mathcal{N}(y; \mu_i, \sigma^2) \quad i = 1, \ldots, 80$$
$$\mu_i = \beta_0 + \texttt{X1}_i\beta_1 + \texttt{X3}_i\beta_2 + \texttt{X4}_i\beta_3$$

   The model can be fit using the `lm()` function.

```
out = lm(formula = y ~ X1 + X3 + X4, data = datax)
```

5. Describe the results of the final model. Use the partial regression plots to simplify the interpretation of the parameters.

```
summary(out)
plot(effects::allEffects(out))
# or, alternatively, car::avPlots(out)
```

   The final model explains about 0.91% of the variance of `y`. Two of the current predictors are positively related, namely `X3` ($\beta_{\texttt{X3}} = 1.438$ and $\sigma_{\beta_{\texttt{X3}}} = 0.039$) and `X3` ($\beta_{\texttt{X4}} = 0.701$ and $\sigma_{\beta_{\texttt{X4}}} = 0.038$), whereas one of them is negatively related to the outcome, i.e. `X1` ($\beta_{\texttt{X1}} = -2.913$, $\sigma_{\beta_{\texttt{X1}}} = 0.040$). The t-statistics associated to the regression coefficients are significant ($\alpha = 0.05$), with $t_{\beta_{\texttt{X1}}}$ showing the largest value.

6. Compute the 98% confidence intervals for the regression coefficients.

```
confint(object = out, level = 0.98)
```

7. Test the hypothesis $H_0 : \beta_1 = -2.50$ against $H_1 : \beta_1 < 2.50$ with $\alpha = 0.085$.

```
refValue = -2.5
beta1_est = -2.9137
beta1_sd = 0.04013
t_beta1 = (beta1_est - refValue)/beta1_sd
p_tb = pt(q = t_beta1, df = 80 - 3 - 1)
print(p_tb)
```

   The probability $\mathbb{P}(T \geq \texttt{t\_beta1}) = 2.165e^{-16}$ (p-value) is lower then the fixed level α. Then, there is evidence against H_0: The estimated β_1 significantly differ from the reference value.

8. Use the fitted model to predict the outcome variable `y` when `X1` $= 15.2$, `X3` $= 10.9$, and `X4` $= -1.2$ (for further details, see `lab2.R` section `A.6`).

```
newdata = data.frame(X1 = 15.2, X3 = 10.9, X4 = -1.2)
out_pred = predict(object = out, newdata = newdata, se.fit = TRUE, level = 0.95,
    interval = "prediction")
print(out_pred)
```

The prediction for the outcome is equal to $\hat{y} = 15.2\hat{\beta}_1 + 10.9\hat{\beta}_2 - 1.2\hat{\beta}_3 = -20.369$ with a 95% CI of $[-22.517, -18.221]$.

1.8

The insurance company *LifeIsLife* provides life and health insurances in higher risks situations. In order to establish how much a customer should pay to get an insurance-based service (i.e., insurance premium), the company would like to use a statistical model that predicts how much a new customer would pay if he or she asked for a life insurance (`insurance_score`). The company interprets this variables in the following way: the higher the score, the higher the insurance premium that should be payed by the customer. From past research, the company knows that the following variables might be considered as predictor of the insurance premium:

- `risk_death`: score of the risk to death accidentally (the higher the score, the higher the risk)
- `risk_layoff`: score of the risk to lose the current job (the higher the score, the higher the risk)
- `risk_sick`: score of the risk to get sick permanently (the higher the score, the higher the risk)
- `purch_power`: score of the capability to buy goods and service (the higher the score, the higher the propensity)
- `savings`: score of the ability to save money during the life (the higher the score, the higher the propensity)
- `age`: age in years

The company has collected a large amount of data over $n = 5000$ customers (see `dataex3_a.csv`). The goal is to select the best predictors of `insurance_score` by means of a Normal linear model. The built model should be then used to predict the insurance premium for new customers.

1. Load the data.

```
datax = read.csv(file = "dataex3_a.csv", header = TRUE)
source("../labs/utilities.R")
```

2. Make a numerical and graphical summary of the relationships among `y` and the predictors. Indicate the support of the outcome variable `y` and the range for the variable `age`. Finally, provide comments about the graphical results.
Note: In order to create the graphical summary, randomly select a subsample of length $n = 500$ from the original dataset by means of the function `random_subsample()`.

```
psych::describe(x = datax)
subdata = random_subsample(X = datax, n = 500)
exploratory_plots(y = subdata[, 1], X = subdata[, 2:7], plot_type = "loess")
```

The support of the outcome variable is the continuous interval $[-68.30, -14.15]$. The variable `age` is between 18 and 36 years old. By visually inspecting the data, a clear negative relationship between `insurance_score` and `age` can be detected. There is also a mild negative relationship between the outcome and `risk_sick` while there are no clear linear pattern between `insurance_score` and the other predictors.

3. Select the best subset of variables which predicts `insurance_score`.

```
leaps_r2(y = datax[, 1], X = datax[, -1])
```

All the variables included in the dataset can be considered as predictors of `insurance_score`.

4. Define and fit a Normal linear model considering the best subset of predictors.

The Normal linear model is as follows:

$$\texttt{insurance\_score}_i \sim \mathcal{N}(y; \mu_i, \sigma^2) \quad i = 1, \ldots, 5000$$
$$\mu_i = \beta_0 + \texttt{risk\_death}_i\beta_1 + \texttt{risk\_sick}_i\beta_2 + \texttt{purch\_power}_i\beta_3 + \texttt{savings}\beta_4 + \texttt{age}\beta_5$$

The model can be fit using the `lm()` function.

```
out = lm(formula = insurance_score ~ ., data = datax)
```

5. Check the Normality of residuals and homoscedasticity for the fitted model. Evaluate eventual non linearities using partial regression plots.

```
plot(performance::check_normality(out))
plot(performance::check_heteroscedasticity(out))
car::avPlots(out)
```

The assumptions of Normality of residuals and homoscedasticity hold for the fitted model. By visually inspecting the partial regression plots, nonlinear patterns do not emerge. However, the plots indicate the presence of potential unusual observations for the five predictors being considered.

6. Run diagnostics to evaluate the presence of unusual observations. Note that for large sample sizes (in this case $n = 5000$), the diagnostics might be too sensitive in discovering unusual observations; in this case, looking for observations showing highest studentized residuals would be the preferred option. Note that, in case of unusual observations, the model should be fit again after having removed the unusual points.

```
check_unusual_observations(fitted_model = out, m = 10)
car::influencePlot(out, bty = "n")
car::leveragePlots(out)
diffbeta_plot(fitted_model = out)
```

The h-values show no clear leverage points for the fitted model. There are no observations clearly classified as outliers (as indicated by the α-values, which are currently higher then the threshold $\alpha^0 = 0.05$). By contrast, there are few observations (i.e., $i = 386$, $i = 4102$) which could potentially be classified as influential observations. This is also suggested by the visual inspection of the influential plot (see function `car::influencePlot()`). The results of the function `diffbeta_plot()` are not definitive with the current value of the reference quantile (the parameter `qs=0.98` in the function) and a higher value should be considered for this parameter (e.g., `qs=0.99998`). However, by crossing the results provided by `check_unusual_observations()`, `influencePlot()`, and `car::leveragePlots()` we might consider to evaluate how large the suggested observations (i.e., $i = 386$, $i = 4102$, $i = 4250$) affect the estimated β's, as follows:

```
X = dfbetas(model = out)
print(X[c(386, 4102, 4250), ])
```

Overall, the suspected observations do not strongly affect the estimated regression coefficients. The single exception is $i = 386$ for the predictor `risk_sick` where the difference is $\delta_{\beta_2} = -0.104$. We decide to remove the point $i = 396$ and to run again the analyses performed so far.

```
datax = datax[-386, ]
out = lm(formula = insurance_score ~ risk_death + risk_sick + purch_power + savings +
    age, data = datax)
```

7. Describe the results of the final model. Use the partial regression plots to simplify the interpretation of the parameters.

```
summary(out)
plot(effects::allEffects(out))
```

The final model explains about 0.75% of the variance of `insurance_score`. As the sample size n is too large, significance tests are no longer of practical utility. The results suggest the existence of a linear relationship between `risk_death` and `score_insurance` ($\hat{\beta}_1 = 0.723$, $\sigma_{\beta_1} = 0.049$), which indicates that the higher the risk of accidental death, the higher the insurance premium. On the contrary, the other predictors negatively affect the outcome variable. In particular, the risk of being sick is negatively associated to `insurance_score` ($\hat{\beta}_2 = -1.418$, $\sigma_{\beta_2} = 0.049$), the purchasing power is negatively related to `insurance_score` ($\hat{\beta}_3 = -0.532$, $\sigma_{\beta_3} = 0.049$), savings negatively affects the outcome variable ($\hat{\beta}_4 = -0.715$, $\sigma_{\beta_4} = 0.049$). Similarly, `insurance_score` decreases as a function of `age` `insurance_score` ($\hat{\beta}_5 = -1.278$, $\sigma_{\beta_5} = 0.049$).

8. Load the dataset `dataex3_b.csv` (it contains data of $n = 10$ new customers). Use the fitted model to predict the insurance premium for the new observations and indicate whether the new customers should pay a higher or lower insurance premium (for further details, see `lab2.R` section `A.6`).

```
datay = read.csv(file = "dataex3_a.csv", header = TRUE)
out_pred = predict(out, newdata = datay, se.fit = TRUE, level = 0.95, interval = "prediction")
print(out_pred)
```

Based on the trained model, we predict that a lower insurance premium should be provided by the new customers.

1.9

The dataset `happy` from the library `faraway` contains data collected from thirty nine students in the MBA class of the University of Chicago. The dataset includes the following variables:

- `happy`: Happiness on a 10 point scale where 10 is most happy
- `money`: Family income in thousands of dollars
- `sex`: 1 = satisfactory sexual activity, 0 = not
- `love`: 1 = lonely, 2 = secure relationships, 3 = deep feeling of belonging and caring
- `work`: 5 point scale where 1 = no job, 3 = ok job, 5 = great job

The variable `happy` is the outcome to be predicted.

1. Load the data. The `love` predictor takes three possible values but mostly takes the value 2 or 3. Create a new predictor called `clove` which takes the value zero if love is 2 or less and 1 otherwise. Use `clove` instead of `love` in the subsequent analyses. Finally, check whether all the categorical variables are correctly codified as `factor` otherwise transform them (use the `factor()` R function).

```r
source("../labs/utilities.R")

datax = faraway::happy
str(datax)

datax$clove = rep(1, NROW(datax))
datax$clove[datax$love <= 2] = 0

datax$sex = factor(x = datax$sex, levels = c(0, 1), labels = c("satisf", "notSatisf"))
datax$clove = as.factor(x = datax$clove)
```

2. Make a numerical and graphical summary of the relationships between the outcome and the predictors. Provide comments about the graphical results.

```r
exploratory_plots(y = datax$happy, X = datax[, -c(1, 4)])
```

By visually inspecting the data, the variable `happy` is positively associated with `money` (although the positive trend might be caused by influential or leverage points) as well as with `work`. Both the subgroups created by the variable `sex` show the same scores of `happy`. With regards to `clove`, participants in the group `clove=1` show a higher level of happy then participants in the complemetary group. However, the latter show higher level of heterogeneity of `happy` scores.

3. Fit a Normal linear model with `happy` as the response and the other four variables as predictors (additive model). Give an interpretation for the meaning of the `clove` coefficient.

```r
mod1 = lm(formula = happy ~ ., data = datax[, -4])
summary(mod1)
```

The predictor `clove` is categorical with two levels and the interpretation of the estimated coefficient $\hat{\beta}_{\texttt{clove}} = 2.296$ ($\hat{\sigma}_{\beta_{\texttt{clove}}} = 0.411$) has to be done by considering the reference level $\hat{\beta}_0$ (the intercept term). In this case, as the model includes more then a single categorical predictor (the variable `sex` is categorical), the intercept includes both the levels `sex:satisf` and `clove:0`. Consequently, $\hat{\beta}_{\texttt{clove}}$ codifies the increment of `happy` from `clove:0` to `clove:1` when `sex:satisf`. Moreover, given the t-statistic at $\alpha = 0.001$ ($t_{\beta_{\texttt{clove}}} = 5.578$), the increment is statistically significant. To see this, write down the mean of the model:

$$\mu_i = \beta_0 + \texttt{money}_i\beta_{\texttt{money}} + z_i^{(1)}\texttt{sex}_i\beta_{\texttt{sex}} + \texttt{work}_i\beta_{\texttt{work}} + z_i^{(2)}\texttt{clove}_i\beta_{\texttt{clove}}$$

with $\mathbf{z}^{(1)} \in \{0,1\}^n$ and $\mathbf{z}^{(2)} \in \{0,1\}^n$ being dummy vectors. For the case $\mathbf{z}^{(1)} = \mathbf{0}$ and $\mathbf{z}^{(2)}\mathbf{0}$ the structural part of the model boils down to

$$\mu_i = \beta_0 + \texttt{money}_i\beta_{\texttt{money}} + \texttt{work}_i\beta_{\texttt{work}}$$

where β_0 include both the levels `sex=satisf` and `sex=0`. Instead, for the case $\mathbf{z}^{(1)} = \mathbf{0}$ and $\mathbf{z}^{(2)} = \mathbf{1}$ the structural part of the model becomes

$$\mu_i = (\beta_0 + \beta_{\texttt{sex}}) + \texttt{money}_i\beta_{\texttt{money}} + \texttt{work}_i\beta_{\texttt{work}}$$

which shows that $\hat{\beta}_{\texttt{sex}} = 2.296$ quantifies the increment from $\hat{\beta}_0 = 3.453$.

4. Produce the graphical plots for the estimated regression coefficients. Compute the marginal estimated effects for the predictors `clove` and `sex`.

```r
plot(effects::allEffects(mod1))

effects::effect(mod = mod1, term = "clove")
effects::effect(mod = mod1, term = "sex")
```

5. Check the Normality of residuals as well as the homoscedasticity of the fitted model. Compute the $\text{diff}_{\hat{\beta}}$ statistic for the predictors of the model.

```
plot(performance::check_normality(mod1))
plot(performance::check_heteroscedasticity(mod1))

diffbeta_plot(fitted_model = mod1)
```

6. Fit a new Normal linear model which include `sex`, `clove`, and the interaction between them. Compute the effects of the fitted model in terms of analysis of variance (use the `anova()` R function). Provide an interpretation of the results.

```
mod2 = lm(formula = happy ~ clove + sex + clove:sex, data = datax)
anova(mod2)
```

The analysis of variance of the fitted model indicates that `clove` contributes to explain the observed variance of `happy` whereas `sex` as well as the interaction term `love:sex` do not. Those variables can be removed from the next analyses.

7. Plot marginal and ineraction effects of the fitted model. Compute the interaction effect numerically.

```
plot(effects::effect(mod2, term = "clove"))
plot(effects::effect(mod2, term = "sex"))
plot(effects::effect(mod2, term = "clove:sex"))

effects::effect(mod = mod2, term = "clove:sex")
```

8. Compute the $(1-\alpha)$ CI for the interaction term of the model ($\alpha = 0.005$).

```
X = confint(mod2, level = (1 - 0.005))
print(X[4, ])
```

1.10

The file `dataex_4.csv` contains data collected from $n = 250$ participants and regards an experiment set up to study how the cognitive fatigue changes as a function of a new neurological drug (`group` $\in \{$`group`, `control`$\}$). The outcome variable has been also evaluated in terms of the particular cognitive task (`task` $\in \{$`A`, `B`, `C`$\}$) participants have been involved in. To control for eventual spurious relationships, two additional variables have been included in the dataset, i.e. `age` and levels of cortisol (`cortisol`). The outcome variable is represented in terms of scores (the larger the score, the higher the cognitive fatigue). The goal is to verify whether the experimental variables affect the outcome by considering the covariates into the analysis.

1. Load the data and check whether all the categorical variables are correctly codified as factor (use `as.factor()` otherwise).

```
source("../labs/utilities.R")
D = read.csv(file = "dataex4.csv")

str(D)
D$task = as.factor(D$task)
D$group = as.factor(D$group)
```

2. Graphically explore the relationships between the response variable and the predictors. Make a graphical representation of the interaction between `group` and `task`.

```
exploratory_plots(y = D$cogn_fatigue, D[, -5], plot_type = "lm")
interaction.plot(response = D$cogn_fatigue, x.factor = D$group, trace.factor = D$task,
    bty = "n", xlab = "group", ylab = "cogn fatigue", trace.label = "task")
```

The graphical analyses suggest that `cogn_fatigue` (i) decreases in the experimental group as opposed to the control group, (ii) decreases in the second task and increases in the third one as opposed to the first task, (iii) mildly increases as a function of age, and (iv) increases (non-linearly) as a function of cortisol. The empirical density shows a skewed response variable. Moreover, the interaction plot shows a kind of interaction between the two experimental variables: `cogn_fatigue` decreases in the experimental group as opposed to the control one just for the first and third task.

3. Define and fit a first Normal linear model which includes all the variables in the dataset additively along with the interaction between `group` and `task`. Next, in order to model the nonlinearity between `cortisol` and `cogn_fatigue`, define and fit three additional Normal linear models where the variable `cortisol` enters the model as quadratic, cubic, or quartic term. Note: to exponentiate a variable in the `formula` parameter, use the syntax `I(x^a)` where `x` is the variable to be transformed whereas `a` is the corresponding exponent.

```
mod1 = lm(data = D, formula = cogn_fatigue ~ group + task + age + cortisol + group:task)

mod2 = lm(data = D, formula = cogn_fatigue ~ group + task + age + I(cortisol^2) +
    group:task)
mod3 = lm(data = D, formula = cogn_fatigue ~ group + task + age + I(cortisol^3) +
    group:task)
mod4 = lm(data = D, formula = cogn_fatigue ~ group + task + age + I(cortisol^4) +
    group:task)
```

4. Use the posterior predictive check method to identify the best of the four fitted models. In addition, compute the AIC index for each of the fitted models and compare them with the results provided by the posterior predictive check.

```
x11()
par(mfrow = c(2, 2))
posterior_pcheck_Normal(fitted_model = mod1, M = 250, new_window = FALSE)
posterior_pcheck_Normal(fitted_model = mod2, M = 250, new_window = FALSE)
posterior_pcheck_Normal(fitted_model = mod3, M = 250, new_window = FALSE)
posterior_pcheck_Normal(fitted_model = mod4, M = 250, new_window = FALSE)

AIC(mod1, mod2, mod3, mod4)
```

The posterior predictive check procedure suggests `mod3` as the best model. The AIC indices are in line with this conclusion.

5. Verify whether the previous conclusion could been also confirmed using the residual analysis. Plot the distributions of the residuals for each fitted model.

```
x11()
par(mfrow = c(2, 2))
hist(residuals(mod1), prob = TRUE, main = "mod1", bty = "n")
lines(density(residuals(mod1)), lwd = 2, col = "firebrick", lty = 1)
hist(residuals(mod2), prob = TRUE, main = "mod2", bty = "n")
lines(density(residuals(mod2)), lwd = 2, col = "firebrick", lty = 1)
hist(residuals(mod3), prob = TRUE, main = "mod3", bty = "n")
lines(density(residuals(mod3)), lwd = 2, col = "firebrick", lty = 1)
hist(residuals(mod4), prob = TRUE, main = "mod4", bty = "n")
lines(density(residuals(mod4)), lwd = 2, col = "firebrick", lty = 1)
```

The analysis of residuals confirms the conclusions based on posterior predictive check and AIC index.

6. Check the homoscedasticity for the chosen model.

```
plot(performance::check_heteroscedasticity(mod3))
```

The graphical analysis reveals a mild heteroscedasticity for the chosen model. Although the test is significant against homoscedasticity, the graphical analysis shows no *really relevant* patterns in the residuals.

7. Globally evaluate the effects of the independent variables on the response variable. Next, comment the regression coefficients for the predictors `group`, `age`, and `cortisol`.

```
anova(mod3)
summary(mod3)
```

The analysis of variance of the model shows that experimental variables (`group` and `task`) as well as the covariates (`cortisol` and `age`) contribute to explain the variance of the response variable. The interaction term is closed to be significant (at $\alpha = 0.05$). The estimated regression coefficients indicate that, regardless to `task`, `cogn_fatigue` significantly increased in the experimental group as opposed to the control group ($\hat{\beta}_{\text{group:exp}} = 2.351$, $\hat{\sigma}_{\beta_{\text{group:exp}}} = 2.069$, $t_{\beta_{\text{group:exp}}} = 2.199$). Similarly, `cogn_fatigue` significantly increased as a function of `age` ($\hat{\beta}_{\text{age}} = 5.085$, $\hat{\sigma}_{\beta_{\text{age}}} = 0.111$, $t_{\beta_{\text{age}}} = 45.522$) and `cortisol` ($\hat{\beta}_{\text{cortisol}} = 87.00$, $\hat{\sigma}_{\beta_{\text{cortisol}}} < 1e-3$, $t_{\beta_{\text{cortisol}}} > 50.00$).

8. Compute and comment the numerical effects of `task` and `group:task` on the response variable.

```
effects::effect(mod = mod3, term = "task")
effects::effect(mod = mod3, term = "group:task")
```

Marginally, the mean value of `cogn_fatigue` increased as `task` goes from `A` to `C`. Similarly, this pattern still remains when the analysis is computed conditioned on `group` (interaction). In particular, for the control group the outcome increases on average across `task` whereas for the experimental group the outcome variable is the same on average for both the first and second tasks.

9. Represent the effects of `task` and `group:task` graphically.

```
plot(effects::effect(mod = mod3, term = "task"))
plot(effects::effect(mod = mod3, term = "group:task"))

interactions::cat_plot(model = mod3, modx = "task", pred = "group", geom = "line")
```

An alternative way to represent interactions graphically can be performed by using the function `cat_plot` from the `interactions` library.

```
interactions::cat_plot(model = mod3, modx = "task", pred = "group", geom = "line")
```

The interaction emerges between `task=A`, `task=B` and `group`.

10. Graphically explore whether the outcome variable varies as a function of `cortisol` across `group`. Note: you can use the function `cat_plot` from the `interactions` library via the following syntax `interact_plot(model = mod3,pred = "cortisol",modx = "group")`. Similarly, plot the outcome as a function of `cortisol` across `group` and `task` (in this case, the parameter `mod2` of the function `interact_plot` should be used).

```
interactions::interact_plot(model = mod3, pred = "cortisol", modx = "group")
interactions::interact_plot(model = mod3, pred = "cortisol", modx = "group", mod2 = "task")
```

11. Compute the $(1-\alpha)\%$ CI for the regression coefficients ($\alpha = 0.05$). Finally, provide final comments on the results of the analysis.

    Confidence intervals for the regression coefficients can be computed as usual.

```
confint(mod3)
```

However, as we have noticed a mild heteroscedasticity in the model, we might compute robust confidence intervals by using the function `lm_robust()` from the `estimatr` library. Overall, the robust confidence intervals largely resemble those computed using the standard method.

```
mod3_rob = estimatr::lm_robust(data = D, formula = cogn_fatigue ~ group + task +
    age + I(cortisol^3) + group:task, se_type = "HC1")
summary(mod3_rob)
```

Overall, regardless to `task`, the experimental group shows a higher level of `cogn_fatigue` when compared to the control group. Marginally, the new drug seems to produce an opposite effect on the cognitive fatigue. However, by looking at the interaction term, the new drug (`group:exp`) significantly decreases `cogn_fatigue` just for participants involved in the second (`task:B`) and third (`task:C`) tasks. As the interaction reveals, while `cogn_fatigue` shows a decreasing patterns for two of the tasks, the outcome increases for participants in the first task (`task:A`). Also, the outcome varies as a function of the covariates `age` - with older participants showing a higher cognitive fatigue - and `cortisol` - with higher levels of cortisol being associated with higher levels of cognitive fatigue. All in all, the results suggests that the new drug reduces the cognitive fatigue just in the case it can be coupled with cognitive tasks of type `B` and `C`.