Linear Models in R (M1–MIDO)

Lab Session 5 — Student Sheet

Henri PANJO

Table of contents

Setup

To keep numbers readable and reproducible, we set display options:

```
options(scipen = 999, digits = 5)
```

We also load the packages used during this session.

> ⚠ Warning
>
> Don't worry if you don't know them all — we'll introduce functions as we need them. Some provide regression tools, others are for data visualization or diagnostics.

```
library(broom)
library(performance)
library(parameters)
library(datawizard)
library(see)
library(effectsize)
library(insight)
library(correlation)
library(modelbased)
library(glue)
library(scales)
library(GGally)
library(ggpubr)
library(car)
library(lmtest)
library(multcomp)
library(rstatix)
library(matrixTests)
library(ggfortify)
library(qqplotr)
library(patchwork)
library(ggrepel)
library(gtsummary)
library(kableExtra)
library(openxlsx)
library(janitor)
library(marginaleffects)
library(pROC)
library(caret)
library(collapse)
library(tidyverse)
```

```
source("helper_functions5.R")
```

Data import

The file `data05.csv` contains data on binge drinking among 2951 adolescents aged 13 to 19 (`age`), as well as the following variables/predictors: an indicator showing whether the teenager has a friend who drinks alcohol (`friendalc`), sensation seeking on a scale from 5 to 20 (`sensation`), watching movies with alcohol on a scale from 0 to 10 (`filmalc`) and parental supervision in category (`parentsurv`).

The variable `binge` is the dependant variable that indicates if the adolescent does binge drinking.

- We import the data

```
base <- read_csv("data05.csv", show_col_types = FALSE) |>
  mutate(friendalc = factor(friendalc, levels = c("No", "Yes"))) |>
  mutate(binge = factor(binge, levels = c("No", "Yes"))) |>
  mutate(binge_bin = 1 * (binge == "Yes"), .after = binge) |>
  mutate(parentsurv = factor(parentsurv, levels = c("Low", "Medium", "High"))) |>
  relabel(
    friendalc = "Has a drinking friend", parentsurv = "Parental supervision",
    binge = "Binge drinking", age = "Age (years)", sensation = "Sensation seeking score",
    filmalc = "Movies with alcohol score"
  )
```

```
base
```

```
# A tibble: 2,951 x 8
      id binge binge_bin friendalc sensation   age filmalc parentsurv
   <dbl> <fct>     <dbl> <fct>         <dbl> <dbl>   <dbl> <fct>
 1     1 No            0 Yes              15    15   1.56  Low
 2     2 No            0 Yes              10    15   0.941 High
 3     3 No            0 No               15    14   0.808 High
 4     4 Yes           1 Yes              18    16   0.892 Low
 5     5 Yes           1 Yes              13    18   1.94  Low
 6     6 Yes           1 Yes              12    18   6.89  High
 7     7 No            0 Yes              16    14   1.31  High
 8     8 Yes           1 Yes              11    17   2.24  Low
 9     9 No            0 Yes              11    15   2.41  High
10    10 No            0 No                5    14   0.290 High
# i 2,941 more rows
```

- Distribution of categorical variables with `tab_freq1()` from `helper_functions5.R`

```
tab_freq1(base, c("binge", "friendalc", "parentsurv"), digits = 1) |>
  kable(align = "l", padding = 2) |>
  row_spec(c(1, 4, 7), bold = TRUE)
```

Variable	Count (n)	Percent (%)
Binge drinking		
No	2173	73.6%
Yes	778	26.4%
Has a drinking friend		
No	699	23.7%
Yes	2252	76.3%
Parental supervision		
Low	843	28.6%
Medium	975	33.0%
High	1133	38.4%

- Summary of continuous variables with `tbl_summary()` from `{gtsummary}`

```
tab_summary <- select(base, age, sensation, filmalc) |>
  tbl_summary(
    type = list(age ~ "continuous2", all_continuous() ~ "continuous2"),
    statistic = all_continuous() ~ c(
      "{mean} ({sd})", "{median} ({p25}, {p75})", "{min}, {max}"
    ),
    digits = ~ 1
  ) |>
  bold_labels() |>
  as_kable_extra(booktabs = TRUE, linesep = "") |>
  kable_styling(position = "center", latex_options = "HOLD_position")
```

```
tab_summary
```

Characteristic	**N = 2,951**
Age (years)	
Mean (SD)	15.7 (1.4)
Median (Q1, Q3)	16.0 (15.0, 17.0)
Min, Max	13.0, 19.0
Sensation seeking score	
Mean (SD)	12.0 (2.9)
Median (Q1, Q3)	12.0 (10.0, 14.0)
Min, Max	5.0, 20.0
Movies with alcohol score	
Mean (SD)	2.8 (1.6)
Median (Q1, Q3)	2.7 (1.6, 3.9)
Min, Max	0.0, 10.0

Question 1: Exploratory analysis of binge drinking

1. Compute the proportion of binge drinking with respect to the variable `friendalc` and `parentsurv`
2. Compute the mean of the continous variables with respect to the binge drinking status

Question 2: Spliting the data

Create a training database `data_train` (approximately 70% of `base`) and a test database `data_test` (approximately 30% of `base`).

Generalized linear model refresher

Model components

1. Random part :
   distribution of the response y, derived from the exponential family
2. Fixed (systematic), linear predictor:
   linear dependence of explanatory variables, $\mathbf{x}^t\boldsymbol{\beta} = \sum_{j=0}^{p} \beta_j x_j$
3. The link function, generally *canonical*:
   function $g(.)$ that links $E(y|\mathbf{x}) = \mu$ to $\mathbf{x}^t\boldsymbol{\beta}$: $g(\mu) = \mathbf{x}^t\boldsymbol{\beta}$

How do I return to the data scale ?

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}^t\boldsymbol{\beta}$$
$$\mu = E(y|\mathbf{x}) = g^{-1}(\mathbf{x}^t\boldsymbol{\beta})$$

$$g^{-1}(.) = \begin{cases} \dfrac{\exp(.)}{1+\exp(.)} & \textbf{if logistic model} \\ \exp(.) & \textbf{is Poisson model} \end{cases}$$

Question 3: Logistic model

- $y =$ `binge_bin` la réponse pour un étudiant (`binge_bin = 0 ou 1`)
- $\mathbb{E}(y \mid \mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x}) = p(\mathbf{x})$

1. Fit the folowing logistic model (`mod1`), interpret the output test the significance of each variables

$$\log\left[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right] = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{sensation} + \beta_3 \texttt{filmalc} + \beta_4 \texttt{parentsurv}_{\texttt{Medium}} + \beta_5 \texttt{parentsurv}_{\texttt{High}} + \beta_6 \texttt{friendalc}_{\texttt{Yes}}$$

Question 4: Prediction grid and model-based probabilities

1. Using the fitted model, construct a new dataset (`grid1`) that contains **all combinations** of :

- `age` taking values from **13 to 19**,
- all observed levels of `parentsurv`,
- all observed levels of `friendalc`,
- `sensation` and `filmalc` at their means

2. Using `grid1`, compute the **expected probability** of `binge_bin = 1` implied by the fitted logistic model, together with **95% confidence intervals**.

3. Produce a figure displaying the model-based predictions with the following characteristics:

- x-axis: `age`,
- y-axis: predicted probability of `binge_bin = 1`,
- color: `friendalc`,
- separate panels (facets) for `parentsurv`,
- include both points and connecting lines.

Question 5: Predicted values on the test data

1. Use `augment()` from `{broom}` to obtain the values predicted by the `mod1` model based on the test data `data_test`. The database obtained with `augment()` will be named `pred_test`

2. Using `ggplot()`, plot the histogram of the predicted values

Question 6: Confusion matrix

Recall the *confusion matrix*

Threshold s	$y = 0$	$y = 1$
$\hat{y} = 0$	TN	FN
$\hat{y} = 1$	FP	TP

- True Positive (TP): Number of observations that are correctly predicted positives.
- True Negative (TN): Number of observations that are correctly predicted negatives.
- False Positive (FP): Number of observations that are incorrectly predicted positives.
- False Negative (FP): Number of observations that are incorrectly predicted negatives.

The counts in the matrix depend on the threshold s used to classify probabilities estimated by the model

Determine the confusion matrix with $s = 0.5$ and $s = 0.4$. Save the matrices in `confusion50`, `confusion40`

Hint: you can use `confusion_matrix()` from `helper_functions5.R`

Question 7: Accuracy

Accuracy (`acc`) measures the proportion of correct predictions out of the total predictions, providing a general sense of model effectiveness

Threshold s	$y = 0$	$y = 1$
$\hat{y} = 0$	TN	FN
$\hat{y} = 1$	FP	TP

$$\texttt{acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

1. Calculate the accuracy for values of s between 0 and 1 in increments of 0.01. Use
2. Plot the accuracy as a function of s

Question 8: Precision, Positive Predictive Value

Precision (`prec`), also known as *Positive Predictive Value* (`ppv`), measures the accuracy of positive predictions made by a classification model. It quantifies how often the model correctly identifies instances of the positive class. Specifically, precision is defined as the ratio of true positives to the sum of true positives and false positives

Threshold s	$y = 0$	$y = 1$
$\hat{y} = 0$	TN	FN
$\hat{y} = 1$	FP	TP

$$\texttt{prec} = \texttt{ppv} = \frac{TP}{TP + FP} = \mathbb{P}(y = 1 \mid \hat{y} = 1)$$

Negative Predictive Value (`npv`), which assesses the model's performance concerning the negative class

$$\texttt{npv} = \frac{TN}{TN + FN} = \mathbb{P}(y = 0 \mid \hat{y} = 0)$$

1. Calculate the *Precision* (`prec`) and `npv` for values of s between 0 and 1 in increments of 0.01.

Question 9: Recall (Sensitivity, True positive rate), Specificity (True Negative Rate)

Recall (`rec`), also known as *Sensitivity* (`sens`) or *True positive rate* (`tpr`), quantifies the ability of a classification model to correctly identify positive instances from the total actual positives in the dataset. It is defined as the ratio of true positives to the total number of actual positives:

$$\texttt{rec} = \texttt{sens} = \texttt{tpr} = \frac{TP}{TP + FN} = \mathbb{P}(\hat{y} = 1 | y = 1)$$

Specificity (`spec`), *True Negative Rate* (`tnr`) measures the proportion of actual negatives that are correctly identified by the model. It assesses the model's ability to avoid false positives

$$\texttt{spec} = \texttt{tnr} = \frac{TN}{TN + FP} = \mathbb{P}(\hat{y} = 0 | y = 0)$$

1. Calculate the `rec` and `spec` for values of s between 0 and 1 in increments of 0.01.
2. On the same graph, plot *Recall* and *Precision* as a function of s.

Question 10: Receiver Operating Characteristic Curve (ROC curve)

1. Plot the ROC curve, which is the graph of `fpr` $=$ `spec` (False Positive Rate) as a function of *recall* (*sensitivity*, *tpr*)

$$\texttt{fpr} = 1 - \texttt{spec} = 1 - \mathbb{P}(\hat{y} = 0 \mid y = 0) = \mathbb{P}(\hat{y} = 1 \mid y = 0) = \frac{FP}{TN + FP}$$

2. Compute the **Area Under the Curve** (AUC) with `performance_roc()` from `{performance}`

Reminder:
The **Area Under the Curve** (AUC) serves as a robust metric for summarizing the performance of a classification model across all possible thresholds. By comparing the AUC values of different models, we can assess their relative strengths and weaknesses.
AUC quantifies the model's overall ability to discriminate between positive and negative classes. AUC values range from 0 to 1, where 0 indicates that all predictions are incorrect, and 1 indicates that all predictions are correct.