Mathematics for Data Scientists – esoph dataset analysis

KIESGEN de RICHTER Stanislas – VAIO Luca

November 17, 2020

Discovering the dataset

Our dataset comes from a case-control study of œsophageal cancer conducted in Ille-et-Vilaine, France. The dataset is composed of three factors: age groups, alcohol consumption and tobacco consumption. The number of people controlled and proven cases of cancer are given for each group combination. The aim of the study is to confirm the correlation between cancers and age, as well as consumption of alcohol and of tobacco.

To begin with, let's explore the esoph dataset. With R, we can display the first 8 lines, and its dimensions.

```
# Show the dataset shuffled, so that we see different age groups
set.seed(1)
random <- sample(nrow(esoph))
shuffled <- esoph[random,]
head(shuffled, 10)
```

```
##    agegp     alcgp    tobgp ncases ncontrols
## 68 65-74     40-79    10-19      3        10
## 39 45-54    80-119 0-9g/day      3        16
## 1  25-34 0-39g/day 0-9g/day      0        40
## 34 45-54 0-39g/day      30+      0         4
## 43 45-54      120+ 0-9g/day      4         4
## 14 25-34      120+    20-29      0         1
## 82   75+     40-79    10-19      1         3
## 59 55-64      120+ 0-9g/day      5        10
## 51 55-64     40-79 0-9g/day      9        40
## 21 35-44     40-79    10-19      3        23
```

```
cat("Number of combinations: ", dim(esoph)[1], "\nNumber of persons controlled: ",
    sum(esoph$ncontrols))
```

```
## Number of combinations:  88
## Number of persons controlled:  975
```

As we can see, the study takes into account the age group of each individual, along with their alcohol and tobacco consumption per day. The age groups start at 25 and are divided by ranges of 10 years until 75+, the alcohol by ranges of 40g/day and the tobacco by ranges of 10g/day.
In the dataset, each combination of the three groups is presented with the corresponding number of controls and proven cases.

R also provides a summary of the dataset. In this summary below, the first three columns only represent the number of lines of `agegp`, `alcgp` and `tobgp`, which are respectively age, alcohol and tobacco. We cannot get any valuable information from this, because these variables are qualitative. The two other variables, however, are quantitative, so we can analyse the given statistics.

```
summary(esoph)
```

```
##    agegp          alcgp          tobgp        ncases          ncontrols
##  25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00
##  35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00
##  45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00
##  55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08
##  65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00
##  75+  :11                                Max.   :17.000   Max.   :60.00
```

The number of cases goes from 0 to 17 by group combination. The maximum seems to be quite far from the median, and even the third quartile, which is 4. This means that there is a peak in one of the combinations. As we see in the last column, the controls are not equally distributed: they vary from 1 to 60 depending on the group combination. It means that the combinations of the three groups (agegp, alcgp, tobgp) are not equally distributed.

```
tapply(esoph$ncontrols, esoph$agegp, sum)
```

```
## 25-34 35-44 45-54 55-64 65-74   75+
##   116   199   213   242   161    44
```

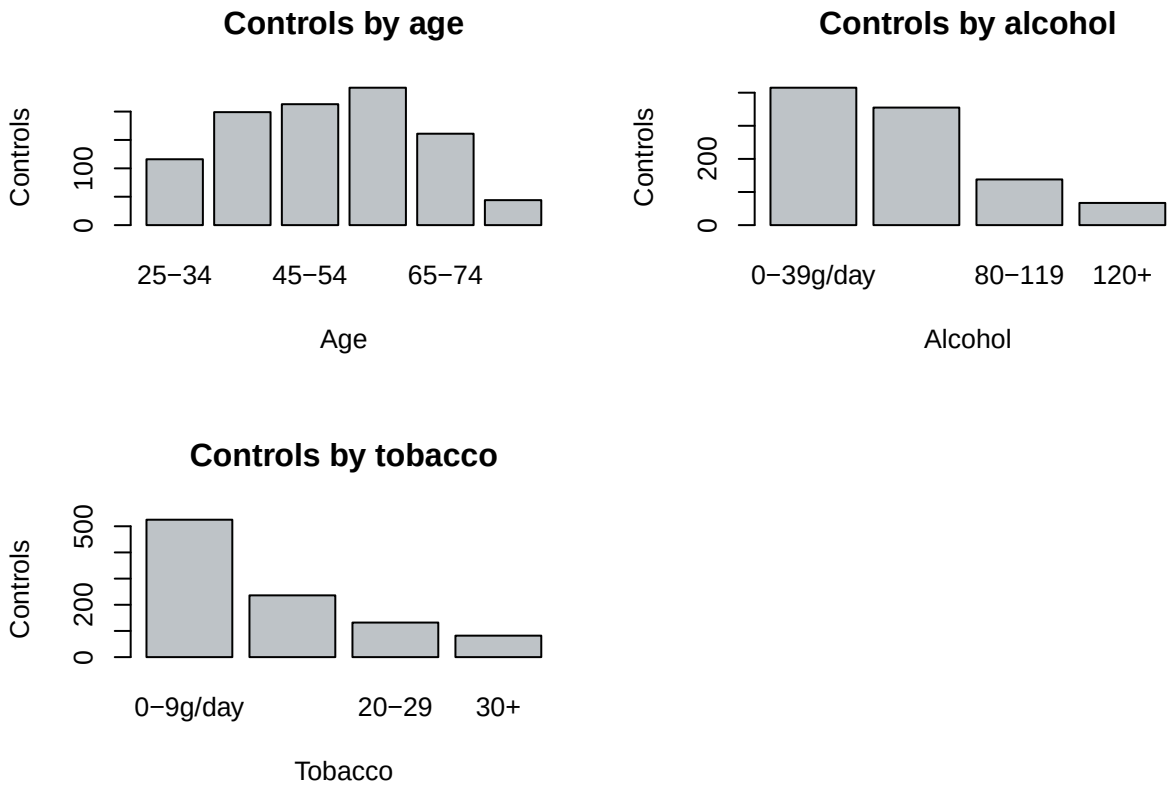
In fact, we can see using the `tapply` R function, which is a `group_by`, that for instance there are 242 controls for the age group `55-64` in the dataset, while there are only 44 for the group `75+`.

On the basis of these observations, we can conclude that we will not be able to base our calculations on the numbers of cases and controls, in particular with regard to the computation of correlation indices between the different variables.

Visualising the data

As said before, the major issue in this dataset is that the quantitative variables are not equally distributed. The following bar plots permit to visualise this well, using as previously the `tapply` function.

```
par(mfrow=c(2, 2))
barplot(tapply(esoph$ncontrols, esoph$agegp, sum), col="#BEC3C7",
        xlab="Age", ylab="Controls", main="Controls by age")
barplot(tapply(esoph$ncontrols, esoph$alcgp, sum), col="#BEC3C7",
        xlab="Alcohol", ylab="Controls", main="Controls by alcohol")
barplot(tapply(esoph$ncontrols, esoph$tobgp, sum), col="#BEC3C7",
        xlab="Tobacco", ylab="Controls", main="Controls by tobacco")
```

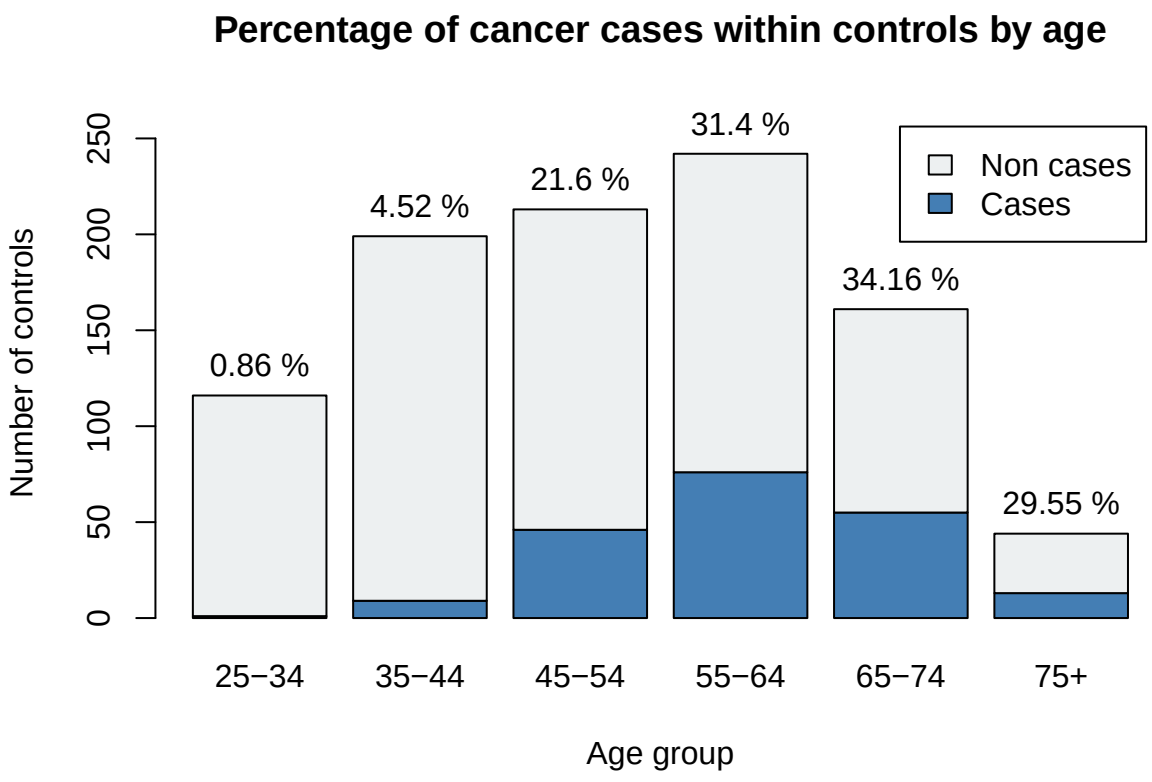


We clearly see that controls not evenly distributed. Note that for alcohol and tobacco, most controlled people present a low consumption, which shows once again that we absolutely cannot overlook the differences in the number of measurements.

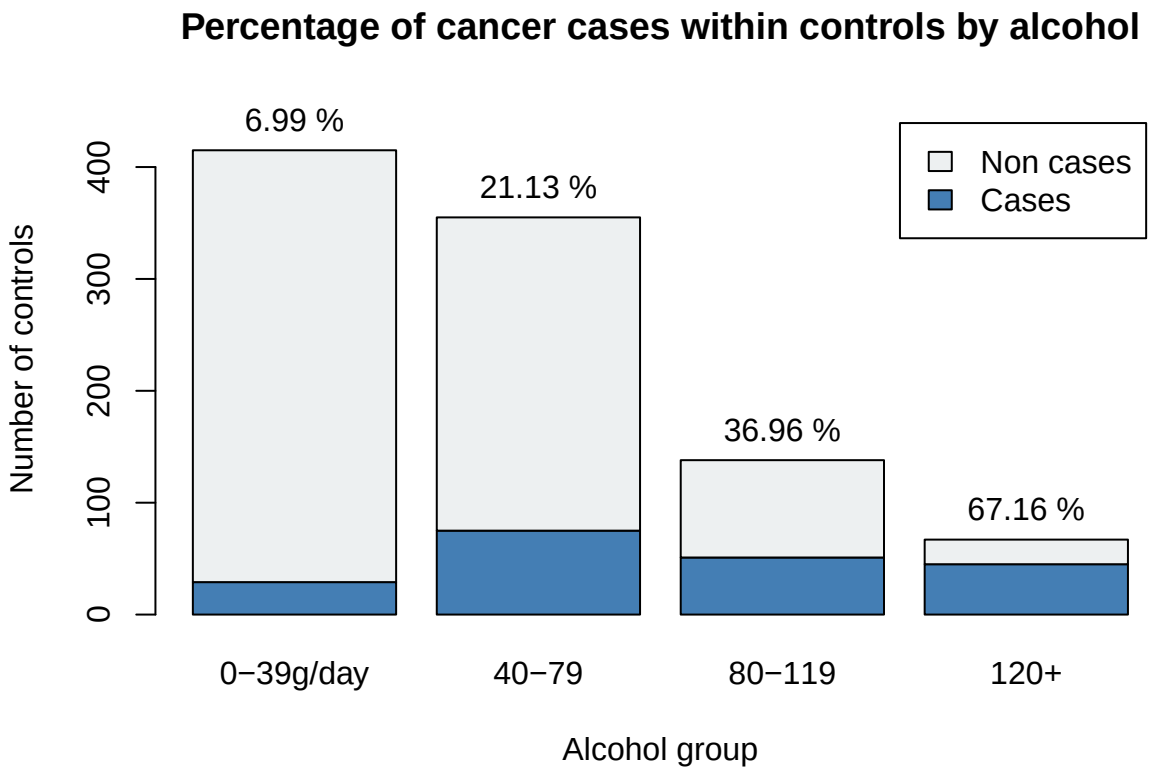
A certainly more reasonable way of analysing this dataset is to base our calculations on the proportion of cancer cases according to the number of controls for each combination of data. A first look at these proportions is given below, by age, alcohol and tobacco.

```
# Age
age.cases <- tapply(esoph$ncases, esoph$agegp, sum)
age.controls <- tapply(esoph$ncontrols, esoph$agegp, sum)
non_age.cases <- age.controls - age.cases
age.proportions <- age.cases / age.controls * 100
```

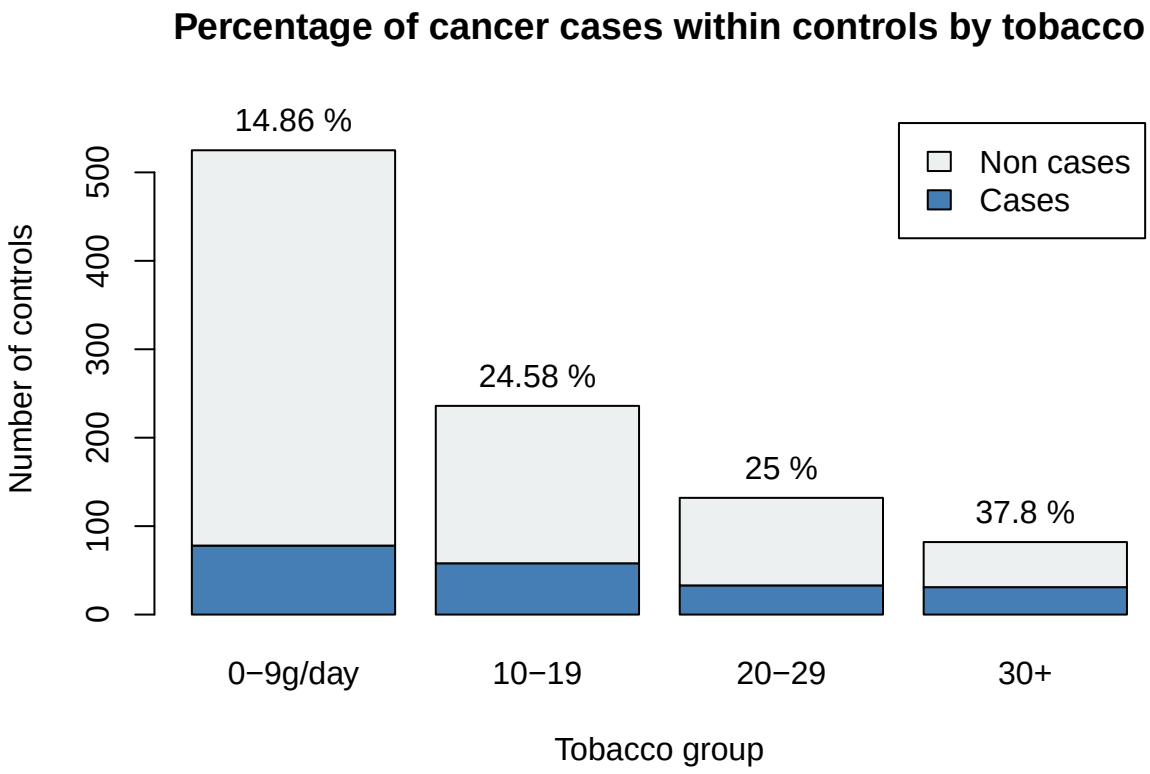
```r
age.ylim <- c(0, 1.1 * max(age.controls))

# Alcohol
alc.cases <- tapply(esoph$ncases, esoph$alcgp, sum)
alc.controls <- tapply(esoph$ncontrols, esoph$alcgp, sum)
non_alc.cases <- alc.controls - alc.cases
alc.proportions <- alc.cases / alc.controls * 100
alc.ylim <- c(0, 1.1 * max(alc.controls))

# Tobacco
tob.cases <- tapply(esoph$ncases, esoph$tobgp, sum)
tob.controls <- tapply(esoph$ncontrols, esoph$tobgp, sum)
non_tob.cases <- tob.controls - tob.cases
tob.proportions <- tob.cases / tob.controls * 100
tob.ylim <- c(0, 1.1 * max(tob.controls))

# Plots
legend <- c("Cases", "Non cases")
colours <- c("#447EB4", "#EDF0F1")
ylab <- "Number of controls"

age.plot <- barplot(rbind(age.cases, non_age.cases), ylim=age.ylim, legend=legend,
                col=colours, xlab="Age group", ylab=ylab,
                main="Percentage of cancer cases within controls by age")
text(x=age.plot, y=age.controls, label=paste(round(age.proportions, 2), "%"), pos=3)
```



```
alc.plot <- barplot(rbind(alc.cases, non_alc.cases), ylim=alc.ylim, legend=legend,
                col=colours, xlab="Alcohol group", ylab=ylab,
                main="Percentage of cancer cases within controls by alcohol")
text(x=alc.plot, y=alc.controls, label=paste(round(alc.proportions, 2), "%"), pos=3)
```



```
tob.plot <- barplot(rbind(tob.cases, non_tob.cases), ylim=tob.ylim, legend=legend,
                col=colours, xlab="Tobacco group", ylab=ylab,
                main="Percentage of cancer cases within controls by tobacco")
text(x=tob.plot, y=tob.controls, label=paste(round(tob.proportions, 2), "%"), pos=3)
```



These three bar plots show, for each group of the concerned factor, the proportion of positive and negative cases within the overall controls. As we saw, and as we still see on these plots, the controls are not equally distributed, but looking at the proportion of cases within those controls is more consistent for our data analysis. In fact, what matters is not how many controls have been done for a given age group for example, but indeed the percentage of positive cancer cases measured.
For instance, the alcohol group `0-39` has 415 controls, but only 6.99% of those controlled people are positive. In parallel, the group `120+` has much less controls (67), but actually a much higher proportion of those are positive, 67.16%, which means that this alcohol group is way more likely to present cancer cases that the previous group.

To finish with, case proportions seem to be more consistent to conduct further analysis, and we will base our computations on this model. The only drawback to this system is that with a different number of controls, the data collected, and therefore the proportions calculated at the same time, will be of very different accuracy. To come back to the previous example, the `120+` group does have 67.16% positive cases, but this proportion could well be revised downwards following further measurements. The `0-39` group has many more measurements and therefore presents more accurate data, which can therefore be relied on more confidently.

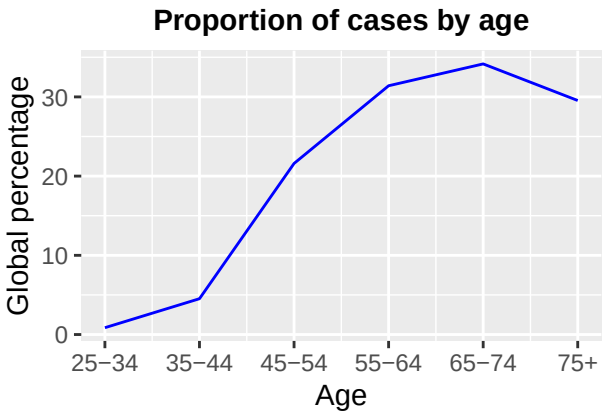
Hypothesis testing

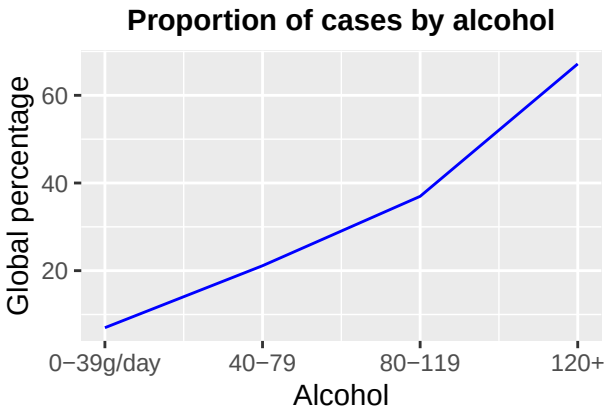
In this section, we want to confirm or not some hypothesis about the dataset. In order to do this, we will define null hypotheses, denoted H_0, and alternative hypotheses, denoted H_a.
The null hypothesis is a default position that there is no relationship between two variables, and the alternative hypothesis is, on the contrary, that there is a specific relationship between those variables.

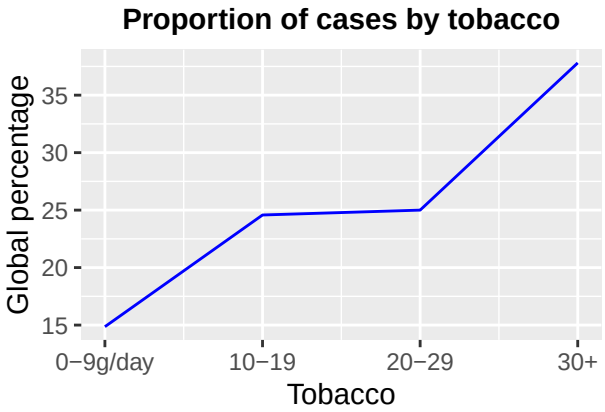
State hypothesis

We have seen that, in order to carry out our analysis, it is preferable for us to work with proportions of positive cases. Now, let's try to plot these proportions graphically by age, alcohol and tobacco groups in order to begin our analysis and state our initial hypotheses.
To do this, let's reuse the global percentages previously calculated for each variable (age, alcohol, tobacco).

```
library(ggplot2)
library(ggpubr)

age.plot <- ggplot(as.data.frame(age.proportions), aes(x=c(0:5), y=age.proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$age)) +
  labs(title="Proportion of cases by age", x="Age", y="Global percentage") +
  theme(plot.title=element_text(size=11, face="bold", hjust=0.5))
alc.plot <- ggplot(as.data.frame(alc.proportions), aes(x=c(0:3), y=alc.proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$alcgp)) +
  labs(title="Proportion of cases by alcohol", x="Alcohol", y="Global percentage") +
  theme(plot.title=element_text(size=11, face="bold", hjust=0.5))
tob.plot <- ggplot(as.data.frame(tob.proportions), aes(x=c(0:3), y=tob.proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$tobgp)) +
  labs(title="Proportion of cases by tobacco", x="Tobacco", y="Global percentage") +
  theme(plot.title=element_text(size=11, face="bold", hjust=0.5))

ggarrange(age.plot, alc.plot, tob.plot)
```







On the graphs, for instance, we clearly see that the proportion of positive cases is higher when the person is old. However, we see that the 75+ group doesn't have so much cases, which can maybe be explained by the fact that the dataset contains far fewer controls for this age group than for other age groups.
But more generally, a clear trend seems to be emerging for all three variables: cancer cases seem to increase with age or consumption. This would mean that the rate of positive cases would be correlated with the variables.

At this stage, we can therefore state three questions:

- Does age favour the development of œsophageal cancer ?
  - H_0: corr $= 0$, i.e. age and cases proportion are not correlated.
  - H_a: corr $\neq 0$, i.e. age and cases proportion are correlated.
- Do alcoholic habits favour the development of œsophageal cancer ?
  - H_0: corr $= 0$, i.e. alcohol and cases proportion are not correlated.
  - H_a: corr $\neq 0$, i.e. alcohol and cases proportion are correlated.
- Do smoking habits favour the development of œsophageal cancer ?
  - H_0: corr $= 0$, i.e. tobacco and cases proportion are not correlated.
  - H_a: corr $\neq 0$, i.e. tobacco and cases proportion are correlated.

In order to answer these questions, we need to test the null hypotheses H_0, which is the default position that there is no relationship between the variables. To do this, we are going to carry out significance tests of the correlations.

Hypothesis verification

First, let's explain briefly what a significance test is about. Eventually, the null hypothesis must be rejected if the p-value determined by the test is lower than the significance level α, generally 0.05. The p-value is the probability of finding the test-statistic value. The latest is the value expected by the null hypothesis, i.e. how closely the distribution matches the null hypothesis. This value is calculated as follows:

$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

Now that the goals are defined, we can start analysing. First of all, we will build a new cleaned dataset called `esoph.prop`, in which all data will be quantitative and where cases will be given by proportion of controls. Therefore, we create a column `cases`, which represents the positive diagnosed cases for each group combination, normalised by number of controls.

```
age <- as.numeric(esoph$agegp)
alcohol <- as.numeric(esoph$alcgp)
tobacco <- as.numeric(esoph$tobgp)
cases <- esoph$ncases / esoph$ncontrols

esoph.prop <- data.frame(age, alcohol, tobacco, cases)
summary(esoph.prop$cases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.2679  0.3468  0.5833  1.0000
```

In fact, we see that this column represents a proportion, from 0 to 1, with a mean of 34.68% positive cases. Note that, unlike the previous graphs here, we have calculated the proportions for each combination, for greater precision, and not the overall proportion for each variable.

Now, we can calculate correlations between all three group variables and cases proportion. Let's first build and display a correlation factors matrix using default Pearson's correlation.

```
esoph.prop.corr <- cor(esoph.prop)
esoph.prop.corr
```

```
##                 age     alcohol     tobacco      cases
## age      1.00000000 -0.01521895 -0.06780840 0.53318708
## alcohol -0.01521895  1.00000000 -0.04896281 0.55068870
## tobacco -0.06780840 -0.04896281  1.00000000 0.08526092
## cases    0.53318708  0.55068870  0.08526092 1.00000000
```

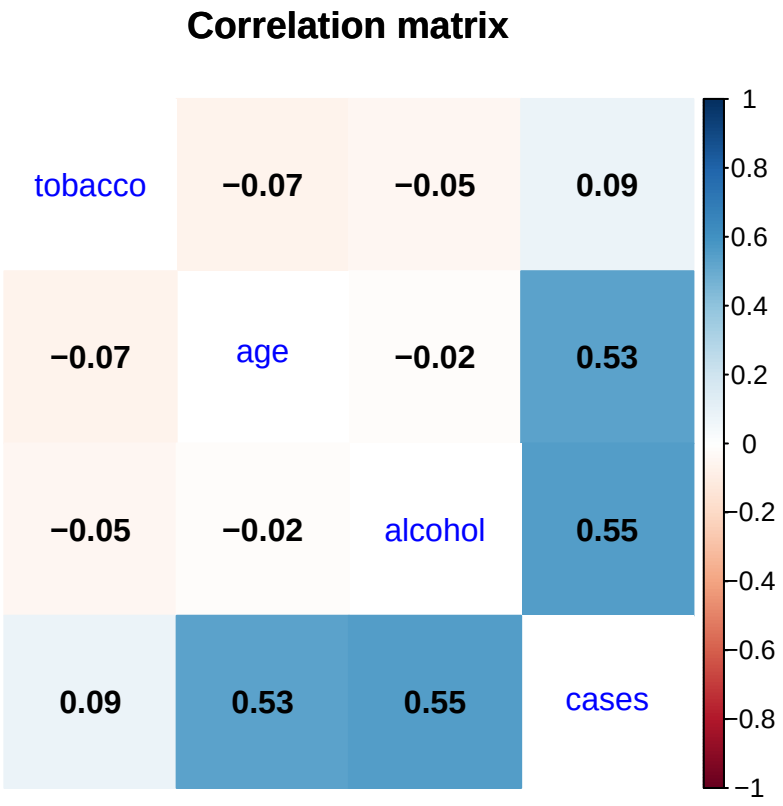
We immediately notice that the diagonal of the matrix is essentially composed of 1, which is quite normal because these are correlations between the same variables.
Also, the correlations between the three qualitative variables of age, alcohol and tobacco have no value here. In fact, this dataset is interesting for its measurements of the number of cancer cases, but the measurements were made in a very homogeneous way between these three variables. In other words, there are supposed to be measures for each combination of these variables, which is why their correlations are almost null.

We can also display the matrix on a clearer way as below, either by masking low values or graphically.

```
cleanTable <- function(table, threshold, subst=NA, rev=FALSE) {
  for (i in seq(1, length(table[,1]))) {
        for (j in seq(1, length(table[1,]))) {
          if (i == j ||
              !is.null(threshold) &&
              xor(rev, table[i, j] > -threshold && table[i, j] < threshold)) {
                table[i, j] <- subst
            }
        }
    }
    return (table)
}

# Masking low values
print(cleanTable(esoph.prop.corr, 0.5), digits=3, na.print=".")
```

```
##           age alcohol tobacco cases
## age         .       .       . 0.533
## alcohol     .       .       . 0.551
## tobacco     .       .       .     .
## cases   0.533   0.551       .     .
```

```
# Graphically
library(corrplot)
esoph.prop.corr.plot <- cleanTable(esoph.prop.corr, NULL)
corrplot.mixed(as.matrix(esoph.prop.corr.plot), lower="color", upper="color",
               order="hclust", title="Correlation matrix", addCoef.col="black",
               tl.col="blue", tl.srt=45, mar=c(0,0,2,5))
```



We do see that age and alcohol are correlated with the proportion of cases, as expected. But surprisingly, it seems that tobacco is not correlated with the proportions of cases, contrary to what seemed to appear previously.
As seen previously, to be able to reject or not the null hypothesis, we must calculate the p-values. For each couple of variables, we can perform a correlation significance test, as follows for age and cases.

```
cor.test(esoph.prop$age, esoph.prop$cases)
```

```
##
##  Pearson's product-moment correlation
##
## data:  esoph.prop$age and esoph.prop$cases
## t = 5.8447, df = 86, p-value = 8.884e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3644429 0.6680292
## sample estimates:
##       cor
## 0.5331871
```

Here, the confidence interval at 95% is $[0.364, 0.668]$. With a correlation factor of 0.533, the confidence interval is fit. In fact, the determined p-value is $8.884 \cdot 10^{-8}$, which is way lower than the significance level $\alpha = 0.05$. Hence, we can reject the null hypothesis that the age is independent of the œsophageal cancer cases. This therefore shows a clear relationship between the age and œsophageal cancers.
Before continuing, we must be careful not to interpret this conclusion in the wrong way. This correlation does not mean that there is a cause and effect relationship, and could even come about by chance.

Let's compute the p-values in a more efficient way (i.e. not one by one), using the `rcorr` function.

```
library(Hmisc)
print(cleanTable(rcorr(as.matrix(esoph.prop))$P, 0.05, rev=TRUE), digits=3, na.print=".")
```

```
##              age  alcohol tobacco    cases
## age            .        .       . 8.88e-08
## alcohol        .        .       . 2.72e-08
## tobacco        .        .       .        .
## cases   8.88e-08 2.72e-08       .        .
```

The result is definitive: the previously identified correlations are significant, and we can reject the null hypotheses that age and alcohol are not correlated with the proportion of cases. However, the tobacco consumption seems not to be sufficiently correlated with cases proportions to reject the associated null hypothesis. We will try to explain this phenomenon later.

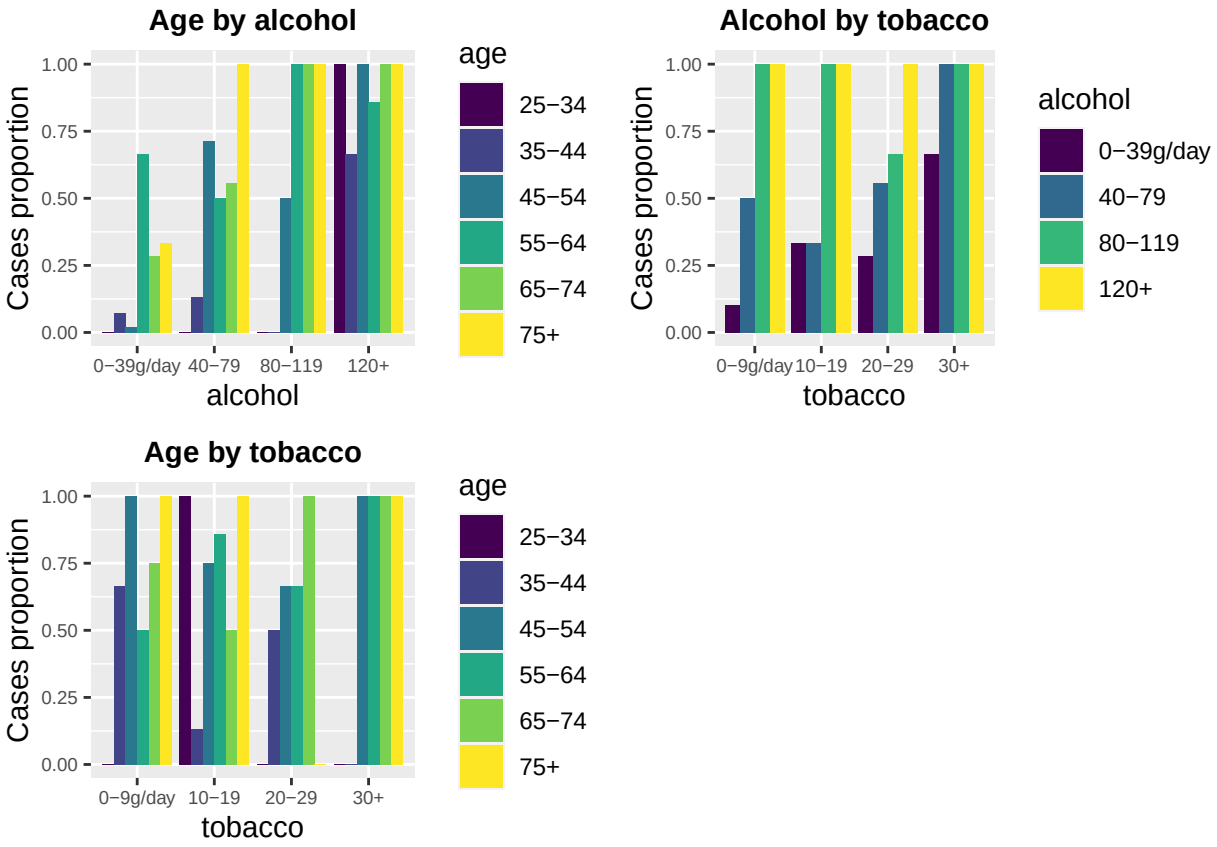
There is one thing to be aware of, however, namely that, as stated above, a correlation between two variables is not sufficient to claim causality. Many factors can influence an observation. Let us imagine, for example, that a sociological study shows that smokers become heavy drinkers with age. In this case, it could very well be that tobacco as well as age influence cases of œsophageal cancer, but not alcohol at all. However, because of this sociological fact, we measure a very strong correlation between alcohol consumption and cases of œsophageal cancer.

After carrying out these last tests, we can confirm that tobacco consumption is not correlated with the proportion of cases of œsophageal cancer. To begin the investigation, let us try to understand how the three variables influence each other.

Influences of the variables on each other

To analyse the influence that these variables have on each other, we will construct three graphs combining the data from one variable according to the categories of another. We could construct all six possible graphs, but these three are sufficient.

```
esoph.prop.qual <- data.frame(esoph$agegp, esoph$alcgp, esoph$tobgp, esoph.prop$cases)
names(esoph.prop.qual) <- c("age", "alcohol", "tobacco", "cases")

age.alc <- ggplot(esoph.prop.qual, aes(alcohol, cases, fill=age)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Age by alcohol", y="Cases proportion") +
  theme(plot.title=element_text(size=11, face="bold", hjust=0.5),
        axis.text=element_text(size=7))
alc.tob <- ggplot(esoph.prop.qual, aes(tobacco, cases, fill=alcohol)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Alcohol by tobacco", y="Cases proportion") +
  theme(plot.title=element_text(size=11, face="bold", hjust=0.5),
        axis.text=element_text(size=7))
age.tob <- ggplot(esoph.prop.qual, aes(tobacco, cases, fill=age)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Age by tobacco", y="Cases proportion") +
  theme(plot.title=element_text(size=11, face="bold", hjust=0.5),
        axis.text=element_text(size=7))

ggarrange(age.alc, alc.tob, age.tob)
```



Let's start with the impact of age as a function of alcohol consumption. We see that from the second alcohol category, `40-79g/day`, there are already 50% or more cases of œsophageal cancer for all age categories except the first two, and that from the third category, `80-119g/day`, the cancer rate is almost always 100% from the age of 55 onwards. Moreover, from an alcohol consumption of 120g per day, almost all cases develop cancer, in any age group or tobacco consumption. This shows the significant impact of heavy alcohol consumption on cancer cases, but also that the age of the person will strongly determine the resistance of his or her body, and in particular that people under 44 years of age suffer much less from it.

Secondly, for the first two age categories, there are very few cases of œsophageal cancer, apart from alcohol consumption exceeding 119g per day. Conversely, as far as tobacco consumption is concerned, cancer cases are very high from the first age category onwards. Indeed, even at low doses of tobacco, and regardless of alcohol consumption or the age of the person, cases of cancer are already high. Moreover, certain data seem to be missing for the first age group.

We may therefore have here a first clue to understanding the non-correlation between tobacco consumption and the proportion of cases. Cases of cancer are said to be immediately high when consumption is low, and therefore to increase less following consumption of tobacco than alcohol, for example.

Building a model

Now that we have shown the correlations between the variables let's try to go further, in order to confirm our hypothesis concerning tobacco consumption.
For that, we can build a model on these data, allowing us to predict a probability of cancer cases according to different criteria. As the data seemed to be quite linear, let's first try to build a linear model and see if it is appropriate.

Simple Linear Regressions

Let's build a linear model on age correlation. The `lm` function takes the y (predictor) and X (variable to model) values, in order to estimate the y-intercept and the slope.

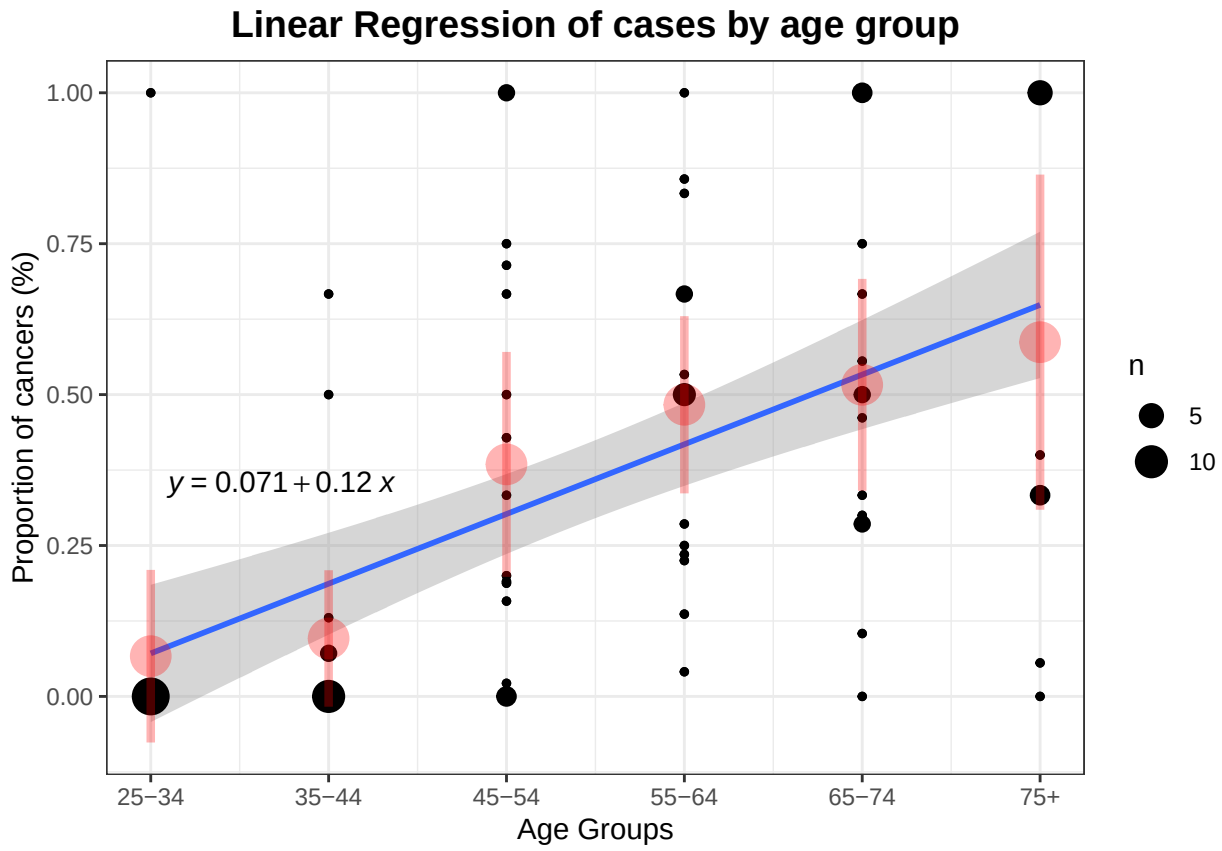
```
y.age <- esoph.prop$age
X <- esoph.prop$cases

age.lm <- lm(X ~ y.age)
summary(age.lm)
```

```
##
## Call:
## lm(formula = X ~ y.age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64861 -0.18672 -0.07125  0.20257  0.92875
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04422    0.07434  -0.595    0.554
## y.age        0.11547    0.01976   5.845 8.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3041 on 86 degrees of freedom
## Multiple R-squared:  0.2843, Adjusted R-squared:  0.276
## F-statistic: 34.16 on 1 and 86 DF,  p-value: 8.884e-08
```

- The residuals are the distance from the observations to the fitted line. A linear model is correct when they are equally distributed around the line, i.e. the statistical values (min, max, etc.) are centered on 0, so especially the quartiles are equidistant and the median is close to 0. Here, it is pretty much the case.
- The estimates are the estimated coefficients of the line.
- The standard errors and the t-values are used to calculate the p-value. As explained before, we want the p-value to be lower than 0.05, meaning that the model will give a reliable estimation of cancer cases proportion.
- The R^2 shows how much the model fits the data, i.e. the proportion of the data that the model could predict. We will discuss this later.

Here, the p-value indicates a significant positive relationship between age group and cancer proportion (p-value $<<$ 0.001), with an increase in cancer proportion of 0.11547 (11.547%) for every unit increase in age group. Then, we can build the models for alcohol and tobacco and visualise the three linear regressions, as shown below.

```
# Age linear regression visualisation

# We remove 1 to start at 0 instead of 1, in order to plot the line from 0.
# As age is a qualitative variable, it does not change the model, only the intercept.
y.age <- y.age - 1

ggplot(esoph.prop, aes(x=y.age, y=X)) + geom_point(size=1) +
  geom_smooth(method="lm") + stat_regline_equation(label.x=0.1, label.y=0.35) +
  stat_sum() + scale_x_continuous(labels=levels(esoph$agegp)) +
  stat_summary(fun.data=mean_cl_normal, col="red", lwd=1.5, alpha=0.3) +
  labs(title="Linear Regression of cases by age group", x="Age Groups",
       y="Proportion of cancers (%)") + theme_bw() +
  theme(plot.title=element_text(size=14, face="bold", hjust=0.5))
```




```
# Alcohol linear regression
y.alc <- esoph.prop$alcohol

alc.lm <- lm(X ~ y.alc)
summary(alc.lm)
```

```
##
## Call:
## lm(formula = X ~ y.alc)
##
## Residuals:
```

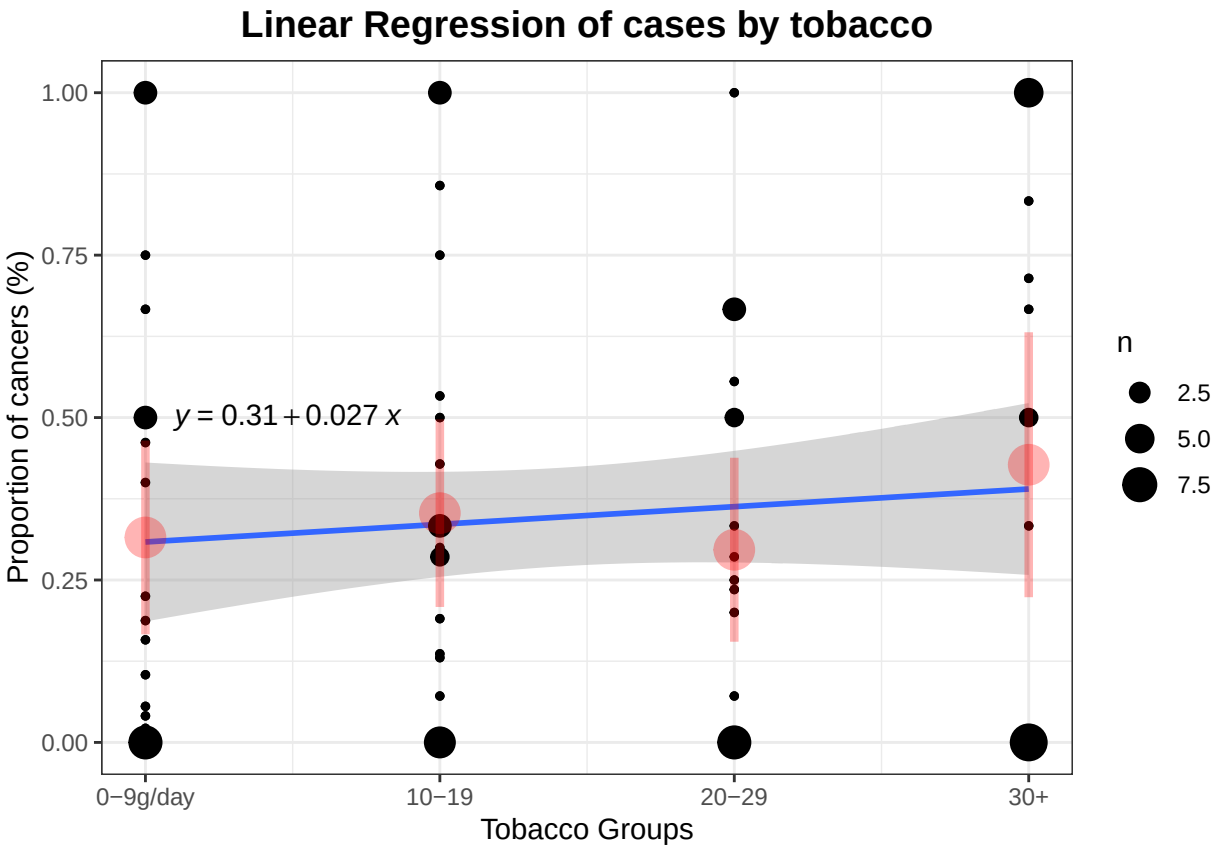
```
##      Min       1Q   Median       3Q      Max
## -0.61750 -0.12231 -0.02626  0.19921  0.73281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08311    0.07720  -1.076    0.285
## y.alc        0.17515    0.02863   6.118 2.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3 on 86 degrees of freedom
## Multiple R-squared:  0.3033, Adjusted R-squared:  0.2952
## F-statistic: 37.43 on 1 and 86 DF,  p-value: 2.716e-08
```

```
y.alc <- y.alc - 1
ggplot(esoph.prop, aes(x=y.alc, y=X)) + geom_point(size=1) +
  geom_smooth(method="lm") + stat_regline_equation(label.x=0.1, label.y=0.35) +
  stat_sum() + scale_x_continuous(labels=levels(esoph$alcgp)) +
  stat_summary(fun.data=mean_cl_normal, col="red", lwd=1.5, alpha=0.3) +
  labs(title="Linear Regression of cases by alcohol", x="Alcohol Groups",
       y="Proportion of cancers (%)") + theme_bw() +
  theme(plot.title=element_text(size=14, face="bold", hjust=0.5))
```

```r
# Tobacco linear regression
y.tob <- esoph.prop$tobacco

tob.lm <- lm(X ~ y.tob)
summary(tob.lm)
```

```
##
## Call:
## lm(formula = X ~ y.tob)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39006 -0.31530 -0.08033  0.21739  0.69150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28131    0.09093   3.094  0.00267 **
## y.tob        0.02719    0.03426   0.794  0.42963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3581 on 86 degrees of freedom
## Multiple R-squared:  0.007269,   Adjusted R-squared:  -0.004274
## F-statistic: 0.6297 on 1 and 86 DF,  p-value: 0.4296
```

```r
y.tob <- y.tob - 1
ggplot(esoph.prop, aes(x=y.tob, y=X)) + geom_point(size=1) +
  geom_smooth(method="lm") + stat_regline_equation(label.x=0.1, label.y=0.5) +
  stat_sum() + scale_x_continuous(labels=levels(esoph$tobgp)) +
  stat_summary(fun.data=mean_cl_normal, col="red", lwd=1.5, alpha=0.3) +
  labs(title="Linear Regression of cases by tobacco", x="Tobacco Groups",
       y="Proportion of cancers (%)") + theme_bw() +
  theme(plot.title=element_text(size=14, face="bold", hjust=0.5))
```



The p-value is very high for the tobacco model, so one more time we can confirm that tobacco is less correlated than age or alcohol with the increase in cancer cases, and that the proportion is already very high for low consumption. In other words, a person who smokes little seems to be almost as likely to develop cancer of the œsophageal as a person who smokes much more, unlike age or alcohol for which the probability increases gradually.

However, we can notice that the error areas are relatively large, this is because the data is not sufficiently linear to better fit the model. Moreover, for all three models, the R^2 value is very low. For instance in the age model, it is of about 0.28, while a correct R^2 should be at least 0.7. This shows indeed that the previous models are not accurate enough.

Thus, let's try to build a multiple linear regression, using all available predictors in the dataset, and see if it is more accurate.

Multiple Linear Regression

The simple linear regressions were not sufficiently accurate, and maybe predicting from all variables would be better. In fact, as we saw before, the qualitative variables age, alcohol and tobacco are not correlated at all, which means they provide significantly different information for the prediction.

With all predictors, the linear regression should look like the following:

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

```
lm <- lm(cases ~ age + alcohol + tobacco, data=esoph.prop)
summary(lm)
```

```
##
## Call:
## lm(formula = cases ~ age + alcohol + tobacco, data = esoph.prop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60571 -0.13109 -0.00011  0.10660  0.67966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.61580    0.09599  -6.415 7.88e-09 ***
## age          0.11955    0.01462   8.177 2.68e-12 ***
## alcohol      0.18017    0.02145   8.400 9.55e-13 ***
## tobacco      0.04796    0.02155   2.226   0.0287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2244 on 84 degrees of freedom
## Multiple R-squared:  0.6191, Adjusted R-squared:  0.6055
## F-statistic: 45.51 on 3 and 84 DF,  p-value: < 2.2e-16
```

In fact, we see here that age and alcohol have very low p-values, so the variables are statistically very significant. This means that using both variables gives a significantly better model than just one of them. However, and once again, tobacco has a quiet lower p-value, meaning that taking into account tobacco will not improve the model as much as the other variables. But still, the p-value is below 0.05, which is good for our model.

Therefore, as expected, the R^2 is higher. With a value of 0.6, the model could potentially predict around 60% of the values correctly. This is fine, but still not great. Thus, let's try to compute a logistic regression on œsophageal cancers proportions.

Logistic regression

Because we want to build a model on categorical variables, it would maybe be better to use a logistic regression. Logistic regression models are constructed using the maximum likelihood method. In other words, the calculated estimates are the values that maximize the likelihood of the observed data. Let's build this logistic model, using all predictors this time.

```
logist <- glm(cases ~ age + alcohol + tobacco, data=esoph.prop, family=binomial)
summary(logist)
```

```
##
## Call:
## glm(formula = cases ~ age + alcohol + tobacco, family = binomial,
##     data = esoph.prop)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4939  -0.3312  -0.1088   0.2759   1.7330
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.5579     1.6910  -4.469 7.84e-06 ***
## age           0.8061     0.2097   3.845 0.000121 ***
## alcohol       1.2032     0.3098   3.883 0.000103 ***
## tobacco       0.3446     0.2622   1.314 0.188775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.609  on 87  degrees of freedom
## Residual deviance: 24.800  on 84  degrees of freedom
## AIC: 72.762
##
## Number of Fisher Scoring iterations: 5
```

We see that both age and alcohol are useful for our model, with a low p-value. However, the p-value for tobacco is very high, so it is not relevant for our modelling.

From these results, let's compute the R^2 value, and its associated p-value. In logistic regressions, McFadden's pseudo-R^2 is calculated as follows:

$R^2 = 1 - \frac{log(L_{full})}{log(L_{null})}$

L_{full} designates the maximized likelihood for the actual model, and L_{null} designates the maximized likelihood for the null model, i.e. without predictors. These values are computed from the deviances given by the R model. The p-value will be estimated from these likelihoods using a χ^2 distribution.

```
lk.full <- -logist$deviance / 2
lk.null <- -logist$null.deviance / 2

logist.pseudo.rsq <- 1 - lk.full / lk.null

logist.p_value <- 1 - pchisq(2 * (lk.full - lk.null),
                             df=(length(logist$coefficients) - 1))

cat("R squared: ", logist.pseudo.rsq, "\np-value: ", logist.p_value)
```

```
## R squared:  0.5974583
## p-value:  5.050488e-08
```

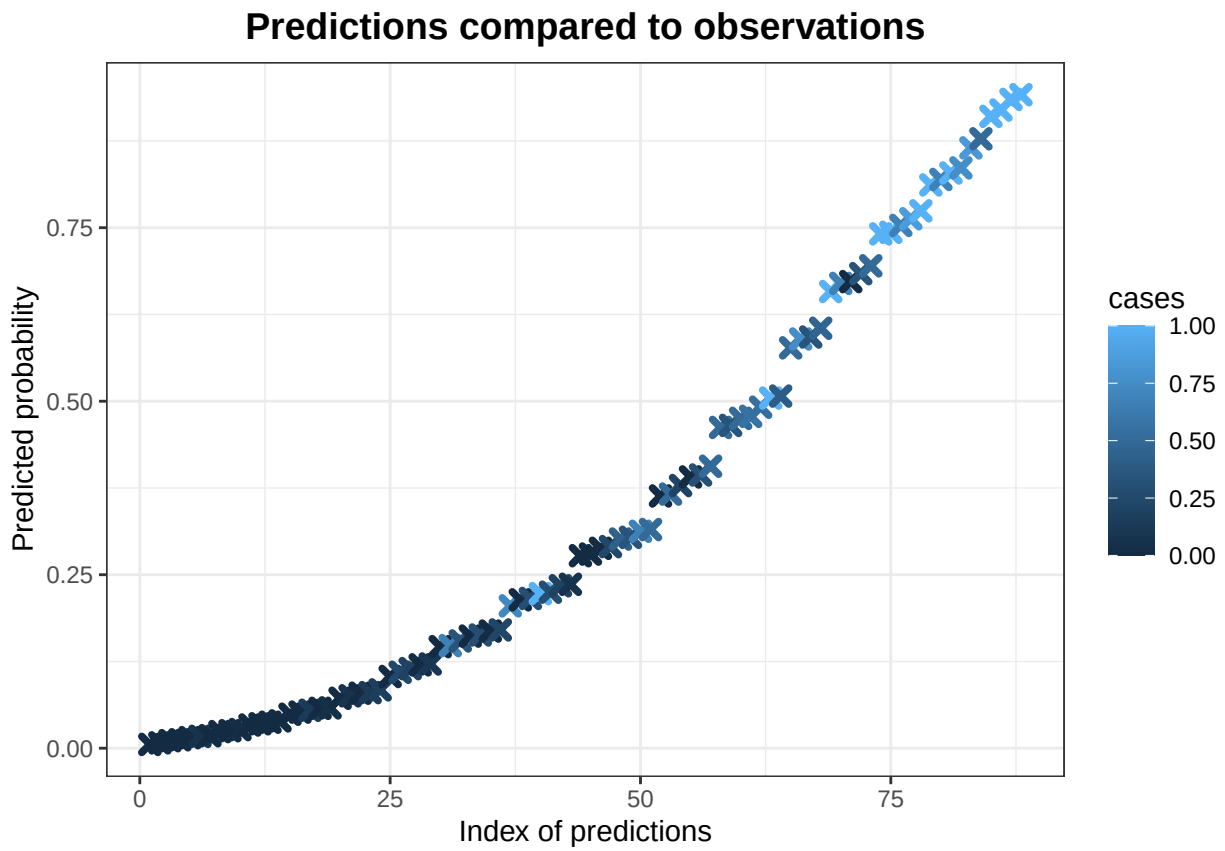
We see that the R^2 is pretty good, 0.6. Unfortunately, it is not much better than using a linear regression. The p-value is way below 0.05, so the statistics here, and so the R^2, are significant, i.e. we can trust the results.

Therefore, these results are trustable and allow us to do some predictions. To do this, as our dataset is very small (only 88 samples), we will not be able to split it into train and test sets. Thus, we will directly use the original train data as the test set, and use the predicted data given with the model to compare.

```
prediction <- data.frame(prob=logist$fitted.values, cases=esoph.prop$cases)

# Add a column with samples indexes in order to plot the results.
prediction <- prediction[order(prediction$prob, decreasing=FALSE),]
prediction$index <- 1:nrow(prediction)

head(prediction)
```

```
##           prob      cases index
## 1  0.005464529 0.00000000     1
## 2  0.007695707 0.00000000     2
## 3  0.010827962 0.00000000     3
## 16 0.012153318 0.00000000     4
## 4  0.015215536 0.00000000     5
## 17 0.017068671 0.07142857     6
```

```
ggplot(data=prediction, aes(x=index, y=prob)) +
    geom_point(aes(color=cases), shape=4, stroke=2) + theme_bw() +
    labs(title="Predictions compared to observations",
         x="Index of predictions", y="Predicted probability") +
  theme(plot.title=element_text(size=14, face="bold", hjust=0.5))
```



This graph shows the predicted probabilities of œsophageal cancers along with the actual cases proportions. We recognize somehow the characteristic S-curve of a logistic regression.
Moreover, we see that samples predicted with a high probability of cancer really have a high proportion of œsophageal cancers. Similarly, we see that samples with a low proportion were predicted with a high probability.

Then, we can measure the accuracy rate of predictions by comparing the predictions with the actual proportions.

```
esoph.prop.precision <- ifelse(esoph.prop$cases > 0.5, 1, 0)
predict.precision <- ifelse(logist$fitted.values > 0.5, 1, 0)

mean(esoph.prop.precision == predict.precision)
```

```
## [1] 0.8409091
```

The model seems to predict correct answers at 84%, which is really great. However, this test is not precise enough because we only test two cases: below and above or equal to 0.5.
For the test to be more consistent, we will melt the values into four categories: below 0.25, 0.5, 0.75 and 1.

```
for (i in esoph.prop$cases) {
  if (i < 0.25) { esoph.prop.precision[length(esoph.prop.precision)+1] <- 0.25 }
  else if (i < 0.5) { esoph.prop.precision[length(esoph.prop.precision)+1] <- 0.5 }
  else if (i < 0.75) { esoph.prop.precision[length(esoph.prop.precision)+1] <- 0.75 }
  else { esoph.prop.precision[length(esoph.prop.precision)+1] <- 1}
}

for (i in logist$fitted.values) {
  if (i < 0.25) { predict.precision[length(predict.precision)+1] <- 0.25 }
  else if (i < 0.5) { predict.precision[length(predict.precision)+1] <- 0.5 }
  else if (i < 0.75) { predict.precision[length(predict.precision)+1] <- 0.75 }
  else { predict.precision[length(predict.precision)+1] <- 1}
}

mean(esoph.prop.precision == predict.precision)
```

```
## [1] 0.7159091
```

With this test, the model gives around 72% correct predictions, which is still very good.

Now, let's predict some probabilities of œsophageal cancers with our model, according to age, alcohol and tobacco consumption.

```
library(glm.predict)

young.clean <- predicts(logist, "1;1;1")$mean
young.consumer <- predicts(logist, "1;4;4")$mean
middle <- predicts(logist, "3;2;2")$mean
old.clean <- predicts(logist, "6;1;1")$mean
old.consumer <- predicts(logist, "6;4;4")$mean

cat("Young and clean: ", round(young.clean*100, 1), "%",
    "\nYoung and heavy consumer: ", round(young.consumer*100, 1), "%",
```

```
    "\nMiddle: ", round(middle*100, 1), "%",
    "\nOld and clean: ", round(old.clean*100, 1), "%",
    "\nOld and heavy consumer: ", round(old.consumer*100, 1), "%")
```

```
## Young and clean:  1.1 %
## Young and heavy consumer:  37.3 %
## Middle:  12.2 %
## Old and clean:  26.1 %
## Old and heavy consumer:  95.6 %
```

In our prediction tests, "clean" means that the person does not drink or smoke, and "heavy consumer" is the opposite. The results are clear: do not drink, nor smoke! According to our model, even being young, so between 25 and 34 years-old, a heavy consuming person has much more chances to develop an œsophageal cancer than an old and clean person.

Conclusion

To conclude, let's quickly summarise what we have seen on this dataset.

First of all, the first observation is that the number of checks is not evenly distributed. In other words, not all combinations of the three variables `agegp`, `alcgp` and `tobgp` were controlled the same number of times. The first step is therefore to create a second cleaned dataset, with no longer the measurements counted, but rather the proportion of positive cases within the number of controls.

Thus, it can be seen that age and alcohol are strongly correlated with the proportion of cancer cases, unlike tobacco consumption. This seems to be explained by the fact that the average number of cases is already very high even at low levels of tobacco consumption.

This is confirmed in the construction of linear models, where tobacco has a less significant importance for the prediction of values. However, it is not to be neglected because the p-value for the multiple linear model, for example, is still sufficiently low.

Finally, our logistic model allows us to predict, according to our estimates, 72% of cancer cases with a rather good confidence. Nevertheless, our models are not perfect. Indeed, we can state that our data set is not sufficiently accurate, containing too few samples. It is therefore possible that our analyses are not quite right, and sometimes even wrong.