Project

KIESGEN de RICHTER Stanislas – VAIO Luca

10/2/2020

Discovering the dataset

Our dataset comes from a case-control study of oesophageal cancer conducted in Ile-et-Vilaine, France. The dataset is composed of a factor of age group, alcohol consumption and tobacco consumption. The number of cases and controls are given for each group. The aim of the study is to confirm the correlation between cancers and age, as well as consummation of tobacco and consummation of alcohol.

Let's first explore the dataset. Here is the beginning, and its summary:

```
head(esoph, 8)
```

```
##   agegp     alcgp    tobgp ncases ncontrols
## 1 25-34 0-39g/day 0-9g/day      0        40
## 2 25-34 0-39g/day    10-19      0        10
## 3 25-34 0-39g/day    20-29      0         6
## 4 25-34 0-39g/day      30+      0         5
## 5 25-34     40-79 0-9g/day      0        27
## 6 25-34     40-79    10-19      0         7
## 7 25-34     40-79    20-29      0         4
## 8 25-34     40-79      30+      0         7
```

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

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

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, 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 cases.

```
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
```

In the summary, the first three columns only represent the number of lines of 'agegp', 'alcgp' and 'tobgp'. We cannot get any valuable information from this, in fact these variables are qualitative:

```
str(esoph)
```

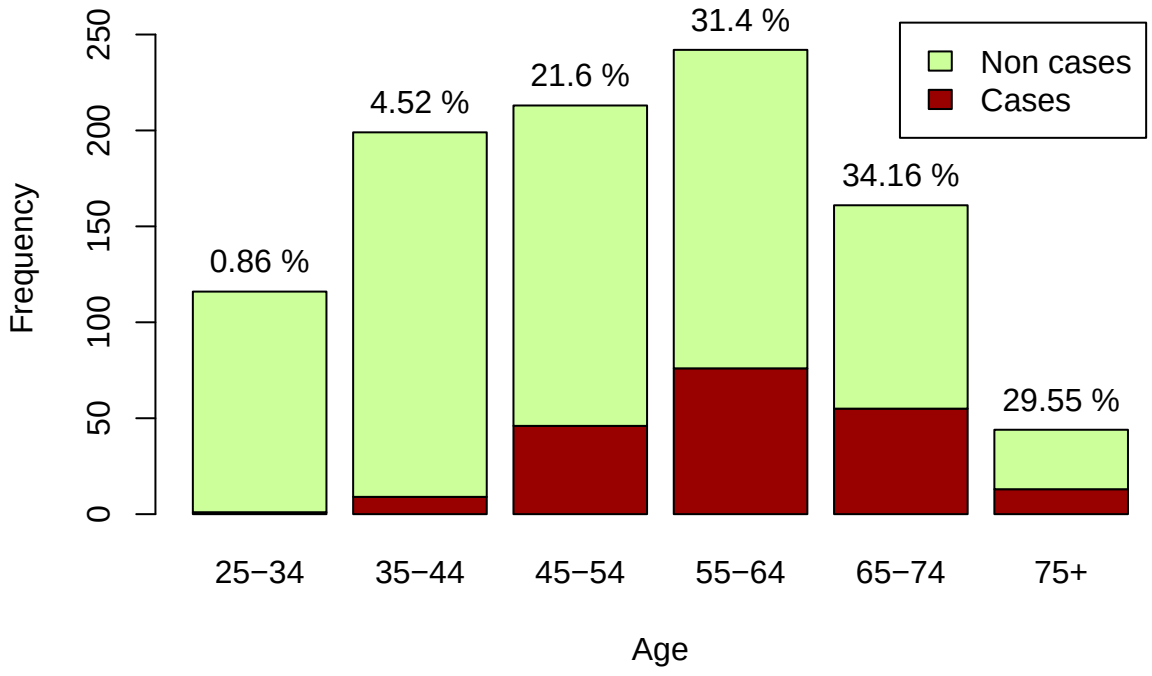
```
## 'data.frame':    88 obs. of  5 variables:
##  $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4 1 2 ...
##  $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 ...
```

The cases go from 0 to 17 by combination. The maximum seems to be quite far from the median, and even the third quartile which is 4, meaning that there is a peak in one of the combinations. As we see in the last column, the controls are not equally distributed: the controls vary from 1 to 60 depending on the group combination. This means that the combinations of the three groups (age, alc, tob) are not equally distributed. Therefore, we will have to compute correlation indexes between the different variables.

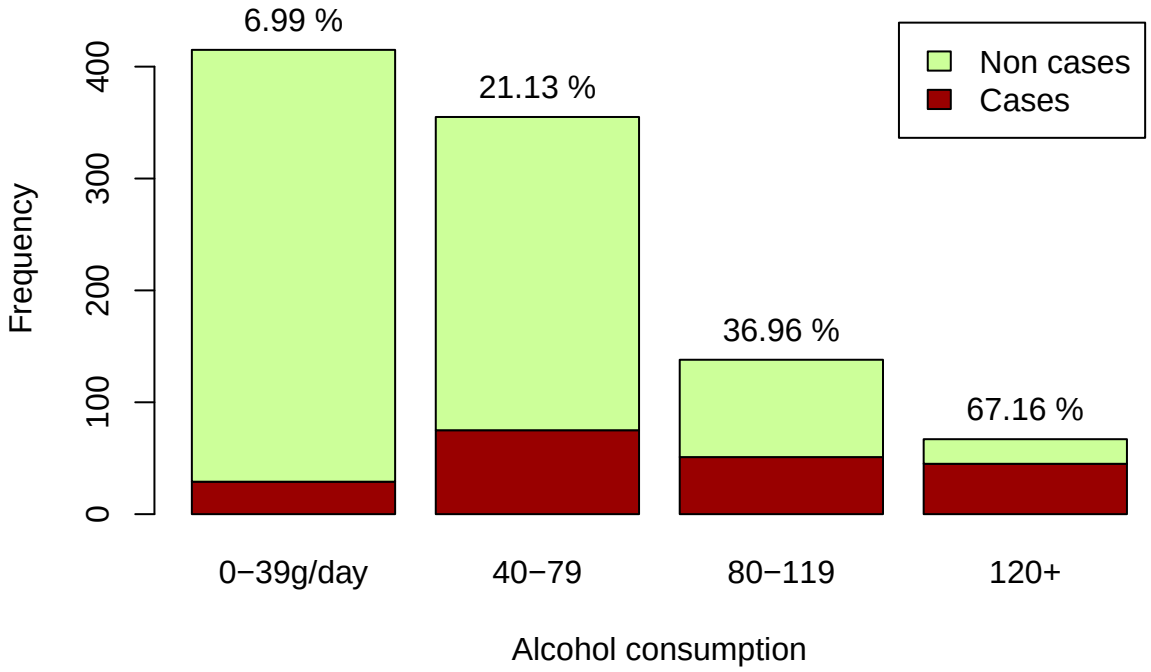
Visualising the data

As said before, one issue is that the controls are not equally distributed, the next bar plot allows to visualize it.

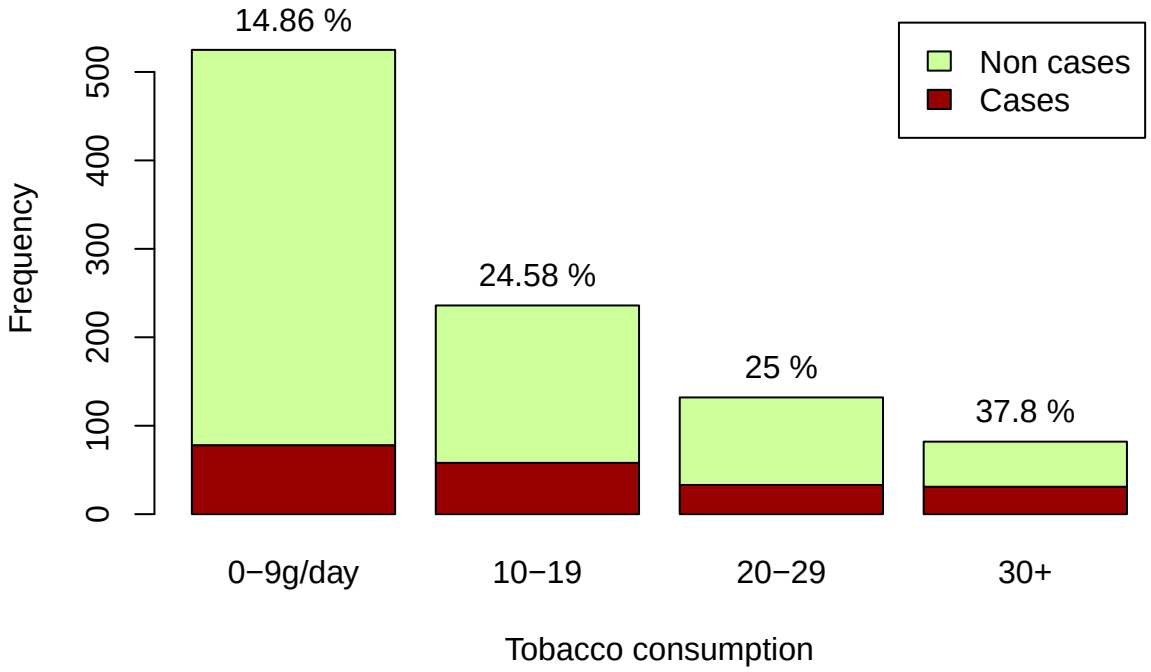
```
cases = tapply(esoph$ncases, esoph$agegp, sum)
controls = tapply(esoph$ncontrols, esoph$agegp, sum)

non_cases = controls - cases
percentages = cases / controls * 100

ylim <- c(0, 1.1*max(controls))
plot <- barplot(rbind(cases, non_cases), ylim=ylim, legend=c("Cases", "Non cases"),
                col=c("#990000", "#CCFF99"), xlab="Age", ylab="Frequency")
text(x=plot, y=controls, label=paste(round(percentages, 2), "%"), pos=3)
```




```
cases = tapply(esoph$ncases, esoph$alcgp, sum)
controls = tapply(esoph$ncontrols, esoph$alcgp, sum)

non_cases = controls - cases
percentages = cases / controls * 100

ylim <- c(0, 1.1*max(controls))
plot <- barplot(rbind(cases, non_cases), ylim=ylim, legend=c("Cases", "Non cases"),
                col=c("#990000", "#CCFF99"), xlab="Alcohol consumption", ylab="Frequency")
text(x=plot, y=controls, label=paste(round(percentages, 2), "%"), pos=3)
```

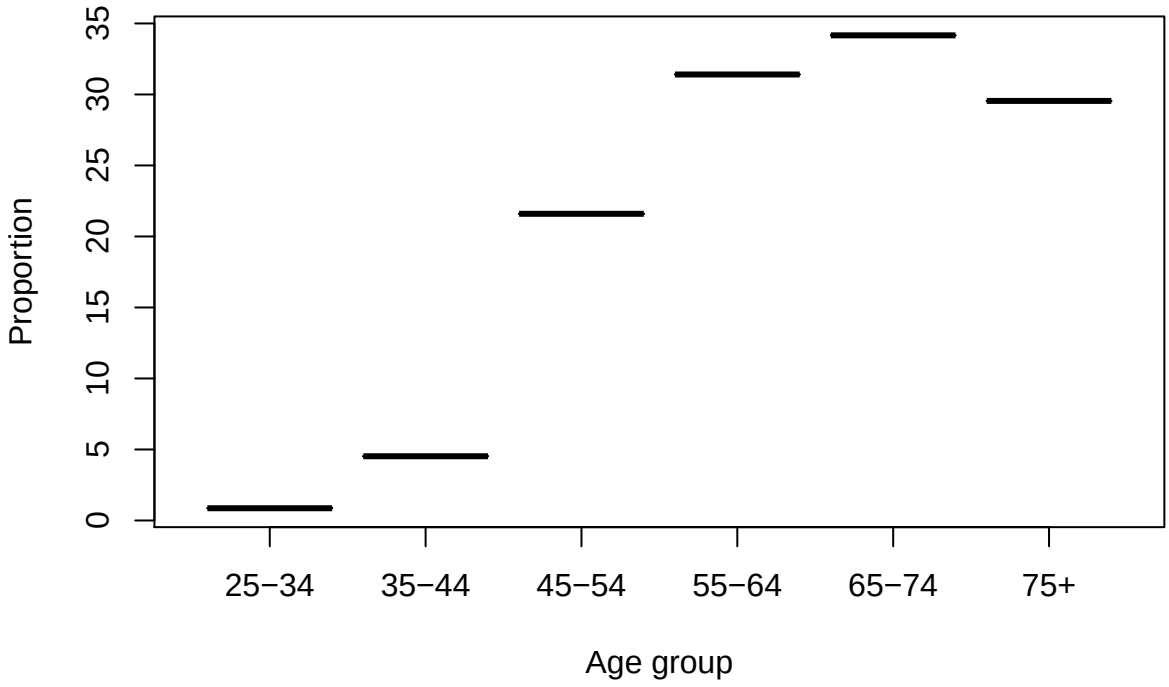



```
cases = tapply(esoph$ncases, esoph$tobgp, sum)
controls = tapply(esoph$ncontrols, esoph$tobgp, sum)

non_cases = controls - cases
percentages = cases / controls * 100

ylim <- c(0, 1.1*max(controls))
plot <- barplot(rbind(cases, non_cases), ylim=ylim, legend=c("Cases", "Non cases"),
                col=c("#990000", "#CCFF99"), xlab="Tobacco consumption", ylab="Frequency")
text(x=plot, y=controls, label=paste(round(percentages, 2), "%"), pos=3)
```



Total proportion of cases by age group. We clearly see that cases are more present from 45 to 75+ years old.

```
total_prop = tapply(esoph$ncases, esoph$agegp, sum) / tapply(esoph$ncontrols, esoph$agegp, sum) * 100

plot(unique(esoph$agegp), total_prop, xlab="Age group", ylab="Proportion")
```



Testing hypothesis

Smoking and cancers correlation

Is there a correlation between smoking habits and oesophageal cancers ? Hypothesis: H0 = tobacco and cases are independent. Ha = tobacco and cases are correlated. Let's use the xtabs function to generate a table showing the number of people with and without cancer corresponding to the smoking categories.

```
tobacco <- xtabs(cbind(ncases, ncontrols) ~ tobgp, data=esoph)
tobacco
```

```
##
## tobgp      ncases ncontrols
##   0-9g/day     78       525
##   10-19        58       236
##   20-29        33       132
##   30+          31        82
```

```
summary(tobacco)
```

```
## Call: xtabs(formula = cbind(ncases, ncontrols) ~ tobgp, data = esoph)
## Number of cases in table: 1175
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 18.363, df = 3, p-value = 0.0003702
```

The capital gain is 0.0003702. This is below the threshold, by default 0.05, so we can reject the H0 hypothesis that smoking and the number of cancers are independent. This therefore shows a relationship between smoking habits and œsophageal cancer.

Age and cancers correlation

Do we get cancer more easily with age (according to alcohol and tobacco consumption) ? We add a column prop_cases, which represents the positive diagnosed cases for each group combination, normalised by number of controls.

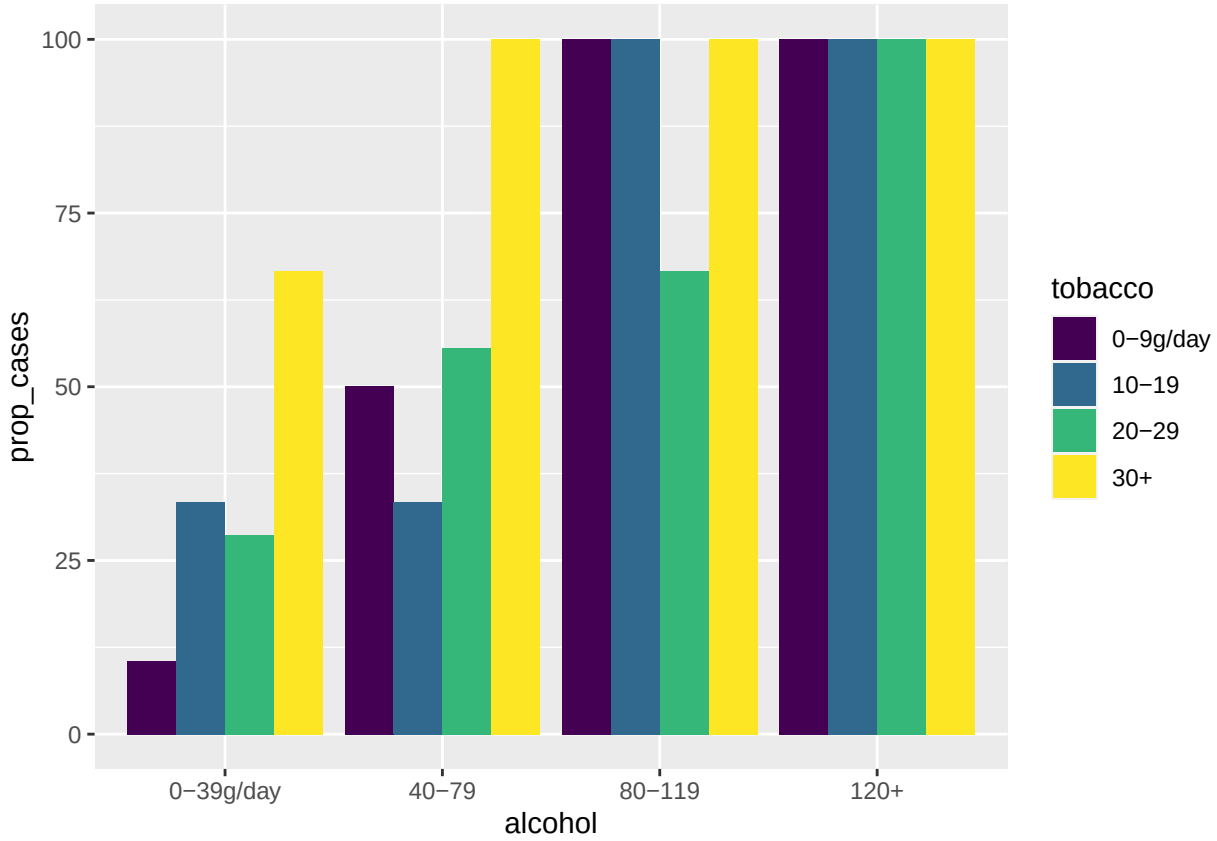
```
age <- esoph$agegp
alcohol <- esoph$alcgp
tobacco <- esoph$tobgp
prop_cases <- esoph$ncases / esoph$ncontrols * 100
esoph_prop <- data.frame(age, alcohol, tobacco, prop_cases)

summary(esoph_prop$prop_cases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   26.79   34.68   58.33  100.00
```

We see that this is in fact a percentage, from 0% to 100%, with a mean of 34.68% positive cases.

First, it can be seen here that alcohol consumption has less of an effect than tobacco on low-dose cancer cases. However, at higher doses, the effects of alcohol catch up with those of tobacco. With alcohol consumption above 80 g/day, tobacco consumption has almost no effect on cancer cases.

```
library(ggplot2)
ggplot(esoph_prop, aes(alcohol, prop_cases, fill=tobacco)) + geom_bar(stat="identity", position="dodge")
```

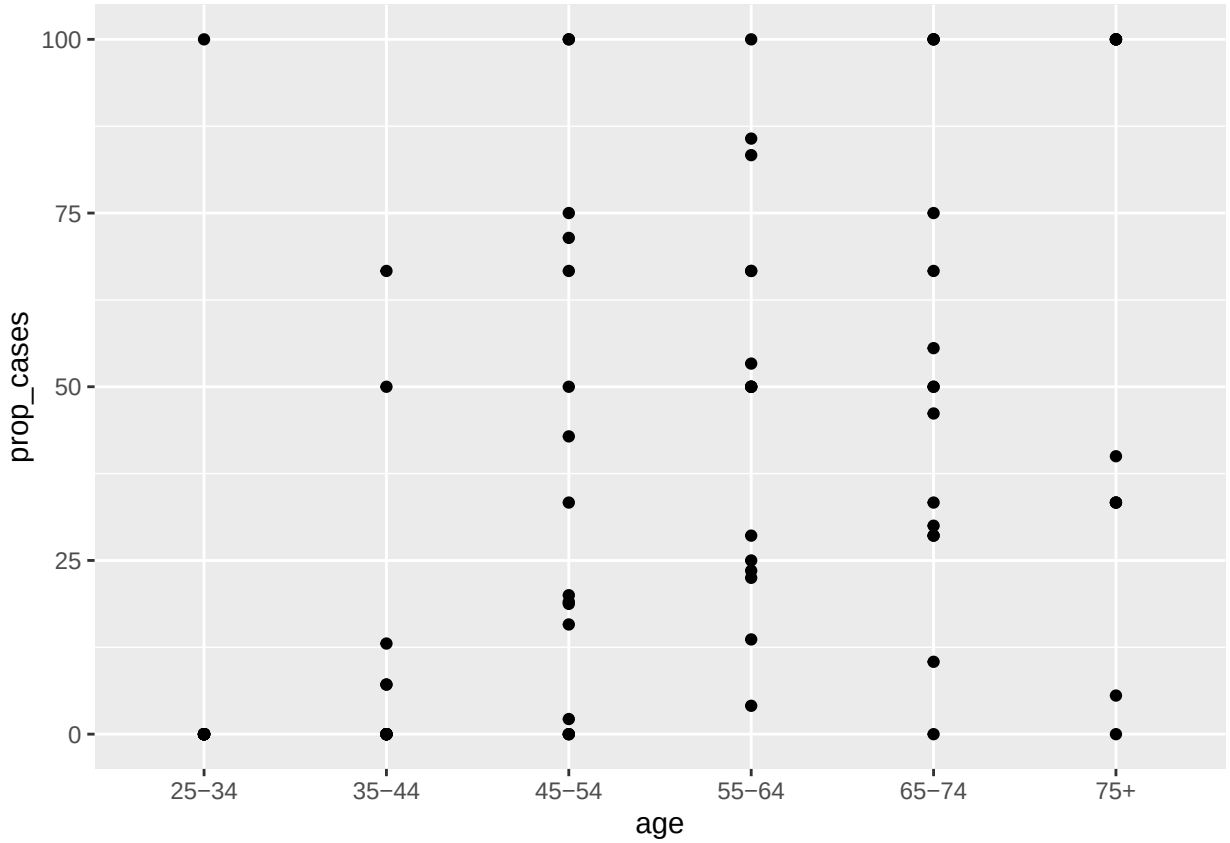


Let's try to build a model.

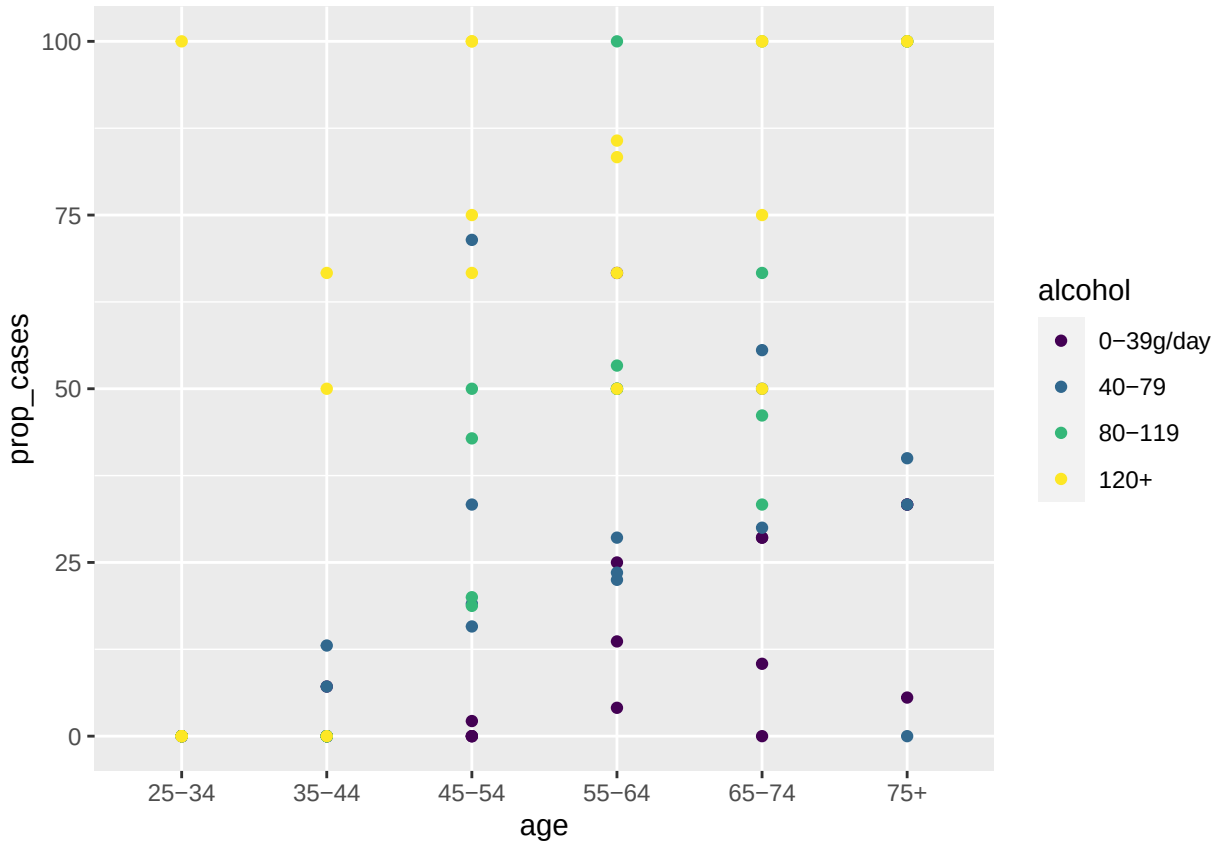
Building a model

Linear Regression

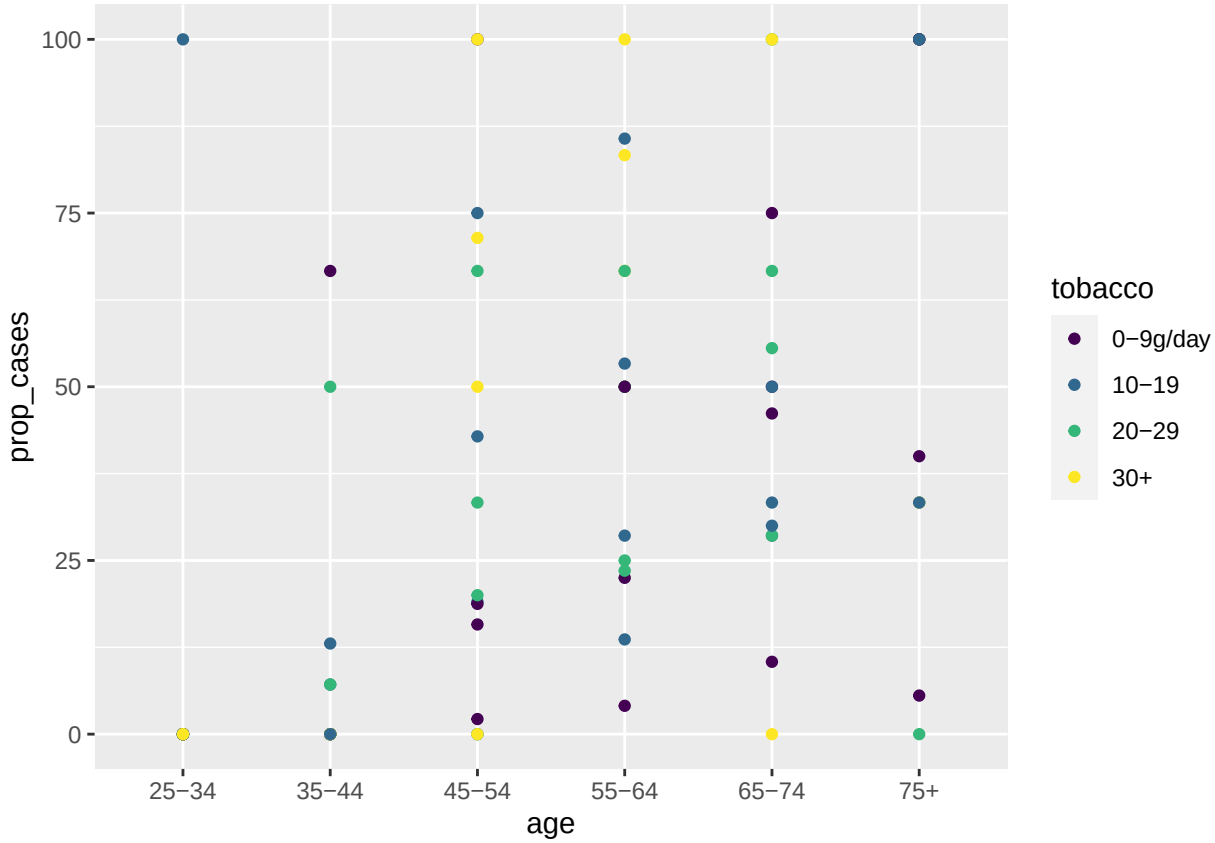
```
# Proportion of cases by age
qplot(age, prop_cases, data=esoph_prop)
```



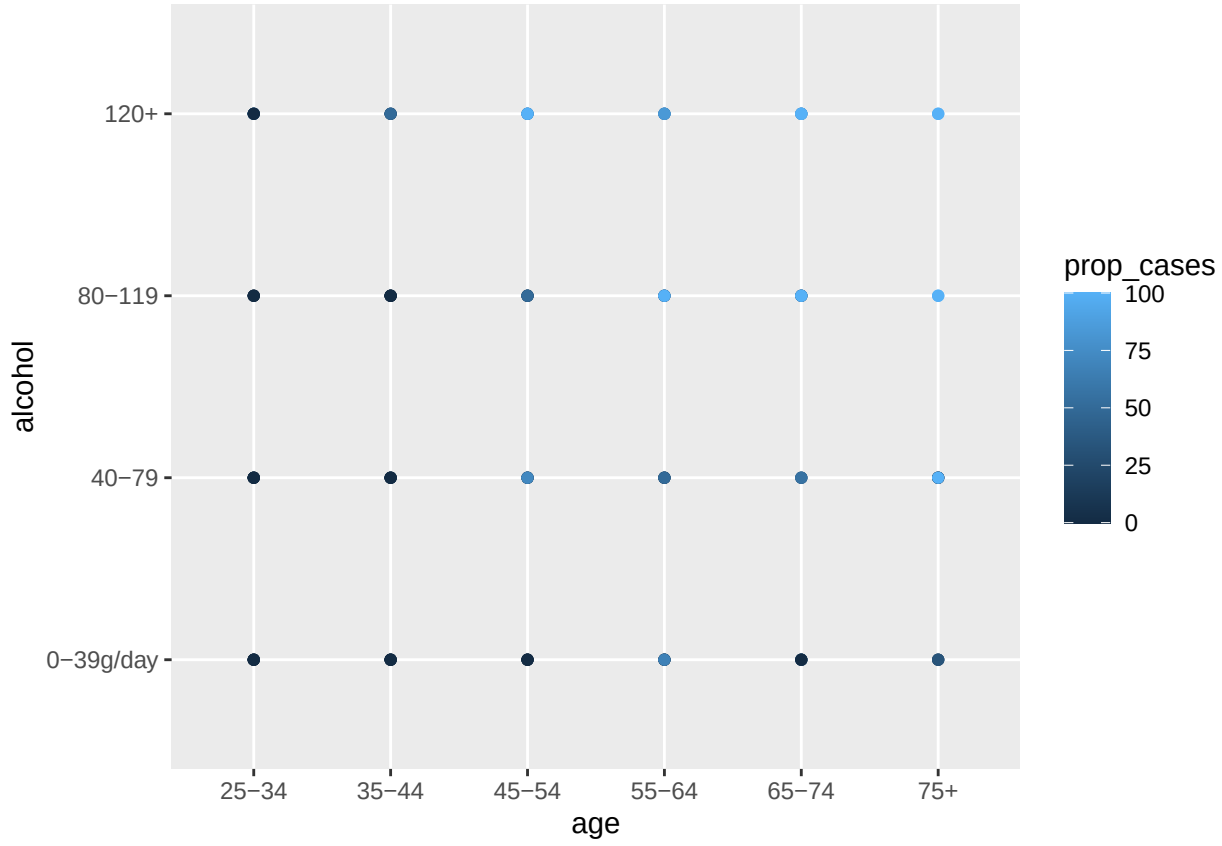

```
# Proportion of cases by age according to alcohol and tobacco consumption
qplot(age, prop_cases, colour=alcohol, data=esoph_prop)
```



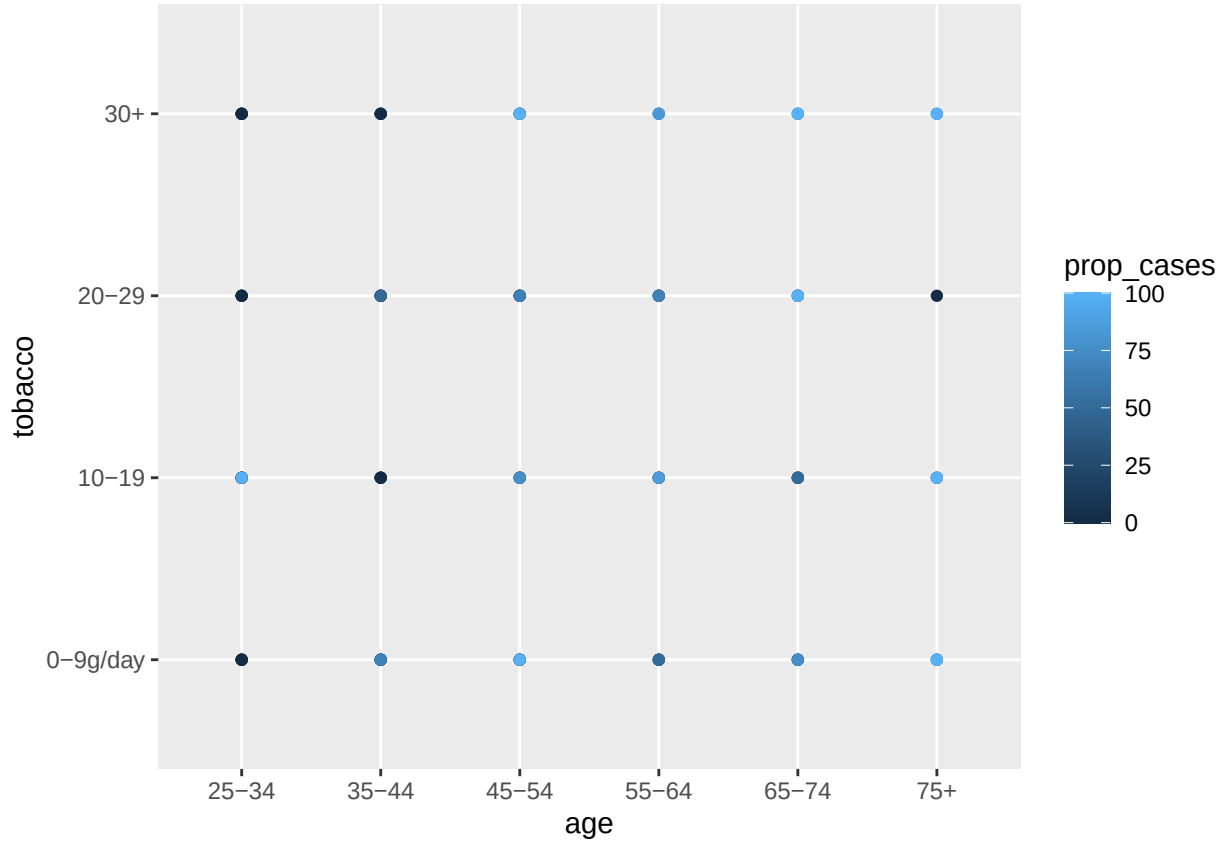

```
qplot(age, prop_cases, colour=tobacco, data=esoph_prop)
```




```
# Proportion of cases by alcohol and tobacco consumption according to age
qplot(age, alcohol, colour=prop_cases, data=esoph_prop)
```




```
qplot(age, tobacco, colour=prop_cases, data=esoph_prop)
```



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 people with cancer die before...

(Model in progress...)

(Attempt of logistic regression)

```
# see http://analyticsdataexploration.com/deviance-and-aic-for-logistic-regression-in-r/
model1 <- glm(cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp, data=esoph, family=binomial())
model2 <- glm(cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp, data=esoph, family=binomial())
summary(model1)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp,
##     family = binomial(), data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6891  -0.5618  -0.2168   0.2314   2.0642
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.77997    0.19796  -8.992  < 2e-16 ***
```

```
## agegp.L      3.00534    0.65215   4.608 4.06e-06 ***
## agegp.Q     -1.33787    0.59111  -2.263  0.02362 *
## agegp.C      0.15307    0.44854   0.341  0.73291
## agegp^4      0.06410    0.30881   0.208  0.83556
## agegp^5     -0.19363    0.19537  -0.991  0.32164
## tobgp.L      0.59448    0.19422   3.061  0.00221 **
## tobgp.Q      0.06537    0.18811   0.347  0.72823
## tobgp.C      0.15679    0.18658   0.840  0.40071
## alcgp.L      1.49185    0.19935   7.484 7.23e-14 ***
## alcgp.Q     -0.22663    0.17952  -1.262  0.20680
## alcgp.C      0.25463    0.15906   1.601  0.10942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  53.973  on 76  degrees of freedom
## AIC: 225.45
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp,
##     family = binomial(), data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8895  -0.5317  -0.2304   0.2704   2.0724
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.75985    0.19822  -8.878  < 2e-16 ***
## agegp.L          2.99646    0.65386   4.583 4.59e-06 ***
## agegp.Q         -1.35008    0.59197  -2.281   0.0226 *
## agegp.C          0.13436    0.45056   0.298   0.7655
## agegp^4          0.07098    0.30974   0.229   0.8187
## agegp^5         -0.21347    0.19627  -1.088   0.2768
## tobgp.L          0.63846    0.19710   3.239   0.0012 **
## tobgp.Q          0.02922    0.19617   0.149   0.8816
## tobgp.C          0.15607    0.19796   0.788   0.4304
## alcgp.L          1.37077    0.21136   6.485 8.85e-11 ***
## alcgp.Q         -0.14913    0.19645  -0.759   0.4478
## alcgp.C          0.22823    0.18203   1.254   0.2099
## tobgp.L:alcgp.L -0.70426    0.41128  -1.712   0.0868 .
## tobgp.Q:alcgp.L  0.12225    0.42044   0.291   0.7712
## tobgp.C:alcgp.L -0.29187    0.42939  -0.680   0.4967
## tobgp.L:alcgp.Q  0.12948    0.38889   0.333   0.7392
## tobgp.Q:alcgp.Q -0.44527    0.39224  -1.135   0.2563
## tobgp.C:alcgp.Q -0.05205    0.39538  -0.132   0.8953
## tobgp.L:alcgp.C -0.16118    0.36697  -0.439   0.6605
```

```
## tobgp.Q:alcgp.C  0.04843    0.36211   0.134   0.8936
## tobgp.C:alcgp.C -0.13905    0.35754  -0.389   0.6973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  47.484  on 67  degrees of freedom
## AIC: 236.96
##
## Number of Fisher Scoring iterations: 6
```

AIC is a maximum likelihood estimate which penalizes to prevent overfitting. It measures flexibility of the models. It is analogous to adjusted R2 in multiple linear regression where it tries to prevent you from including irrelevant predictor variables. Lower AIC of model is better than the model having higher AIC.

Other

75+ less high consumption

```
table(esoph$agegp, esoph$alcgp)
```

```
##
##         0-39g/day 40-79 80-119 120+
##   25-34         4     4      3    4
##   35-44         4     4      4    3
##   45-54         4     4      4    4
##   55-64         4     4      4    4
##   65-74         4     3      4    4
##   75+           3     4      2    2
```

Average number of cases per age and alcohol:

```
tapply(esoph$ncases, list(esoph$agegp, esoph$alcgp), mean)
```

```
##       0-39g/day    40-79 80-119     120+
## 25-34  0.000000 0.000000   0.00 0.250000
## 35-44  0.250000 1.000000   0.00 1.333333
## 45-54  0.250000 5.000000   3.00 3.250000
## 55-64  3.000000 5.500000   6.00 4.500000
## 65-74  2.750000 8.333333   3.25 1.500000
## 75+    1.333333 1.000000   1.00 1.500000
```