Generalised Linear Models – Lab 1

This lab will give you an opportunity to fit models for binary and binomial responses.

1. Class survey data

We will use the data from the class survey for the examples below. Let's begin by looking at the first few rows of the dataset:

```
cs <- read.csv(url("http://www.stats.gla.ac.uk/~tereza/rp/GLMclasssurvey202122.csv"))
nrow(cs)
```

```
## [1] 211
```

```
head(cs)
```

```
##       Year Gender Age EyeColour              GrewUpIn Siblings Pets Speed Shoes
## 1 Previous Female  43     Brown Small town/rural area        0    0     0    15
## 2 Previous   Male  22      Blue Small town/rural area        1    2   110     3
## 3 Previous   Male  21     Black Small town/rural area        3    0     0    30
## 4 Previous   Male  21     Brown Small town/rural area        2    0   160     4
## 5 Previous Female  20      Blue Small town/rural area        1    0    NA    NA
## 6 Previous   Male  20      Blue Small town/rural area        2    0    90     3
##                Coffee                 Tea Astrology          Dress
## 1           Every day A few times a month       Yes White and gold
## 2 A few times a month               Never        No White and gold
## 3  A few times a year               Never        No White and gold
## 4               Never A few times a month       Yes Black and blue
## 5               Never               Never        No Black and blue
## 6               Never               Never       Yes Black and blue
##            Jacket   Hear
## 1 white and green  Yanny
## 2 Black and brown Laurel
## 3 Black and brown Laurel
## 4  White and blue   Both
## 5 Black and brown
## 6  White and blue Laurel
```

We have binary, categorical and continuous variables. For this lab, we will work with binary variables, such as whether or not a student drives. Please note that this year's responses correspond to `Year=2022`, while responses from previous years are labelled as `Year=Previous`. You can use do the analyses for this lab with either just the 2022 data or the entire dataset, including previous years' responses, as is the case below.

Driving

The survey question was actually about how fast you've ever driven a car, but since people were asked to answer 0 if they don't drive, we can create a binary variable, `Drive`, which will take the value 1 for those who drive, and 0 for those who don't. There may be some NAs too.

```
cs$Drive <- NA
cs$Drive[cs$Speed==0] <- 0
```

```
cs$Drive[cs$Speed>0] <- 1
cs$Drive <- factor(cs$Drive, labels=c("No", "Yes"))
summary(cs$Drive)
```

```
##   No  Yes NA's
##   56  151    4
```

```
table(cs$Drive)
```

```
##
##  No Yes
##  56 151
```

We may also wish to combine levels of other factors which will be considered as predictors, e.g. where people grew up:

```
cs$GrewUpIn <- factor(cs$GrewUpIn)
levels(cs$GrewUpIn)
```

```
## [1] ""                         "Big city"
## [3] "Large Town"               "Small town/rural area"
## [5] "Suburbs"                  "Suburbs outside big city"
## [7] "Village"
```

```
cs$GrewUpB <- "Other"
cs$GrewUpB[cs$GrewUpIn=="Big city"] <- "Big city"
cs$GrewUpB[cs$GrewUpIn=="Large Town"] <- "Big city"
cs$GrewUpB <- factor(cs$GrewUpB)
levels(cs$GrewUpB)
```

```
## [1] "Big city" "Other"
```

```
table(cs$GrewUpB)
```

```
##
## Big city    Other
##       94      117
```

Now let's look at some exploratory plots. Bar charts are helpful for exploring associations between the binary response and categorical variables such as gender.

First let's take a look at the gender variable:

```
cs$Gender <- factor(cs$Gender)
levels(cs$Gender)
```

```
## [1] ""           "Female"     "Male"       "non binary"
```

There are two responses that could be combined into an "Other" category. As this number is small, it may make sense to exclude values that are not labelled as "Male" or "Female" from models that include `Gender` as an explanatory variable.

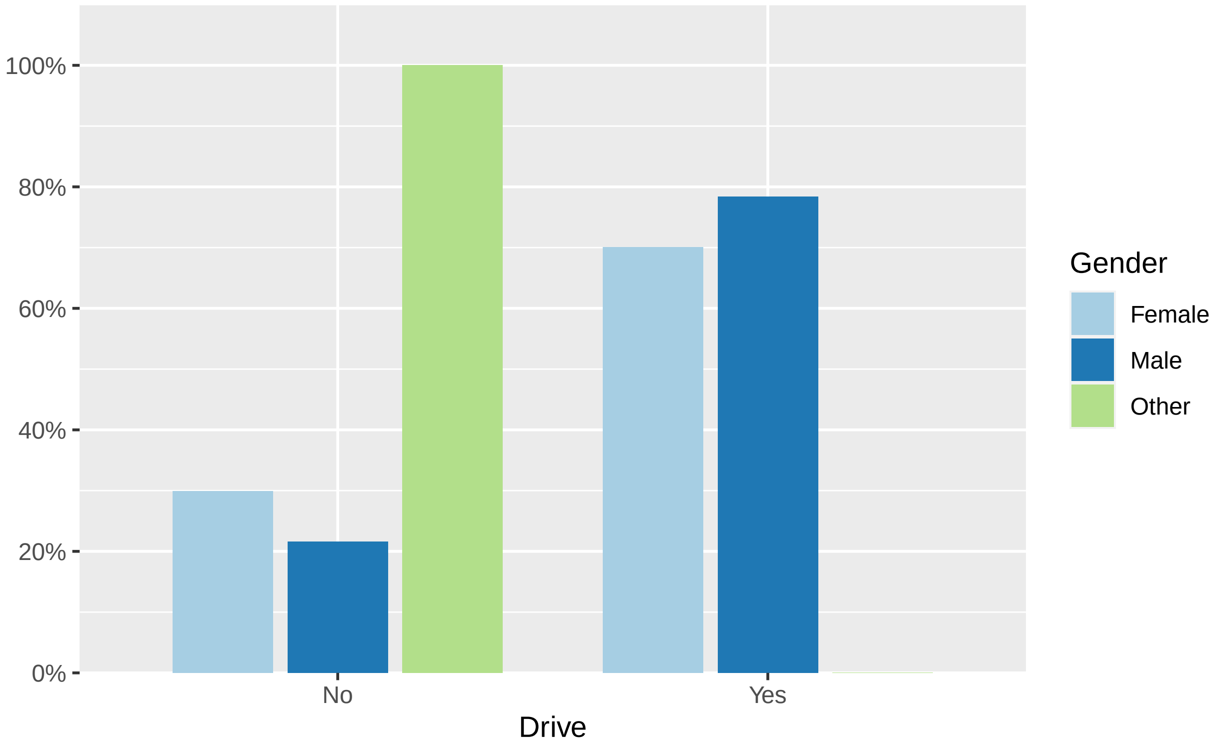
```
cs$Gender2 <- NA
cs$Gender2[cs$Gender=="Female"] <- "Female"
cs$Gender2[cs$Gender=="Male"] <- "Male"
cs$Gender2[cs$Gender==""] <- "Other"
cs$Gender2[cs$Gender=="non binary"] <- "Other"
table(cs$Gender2)
```

```
##
```

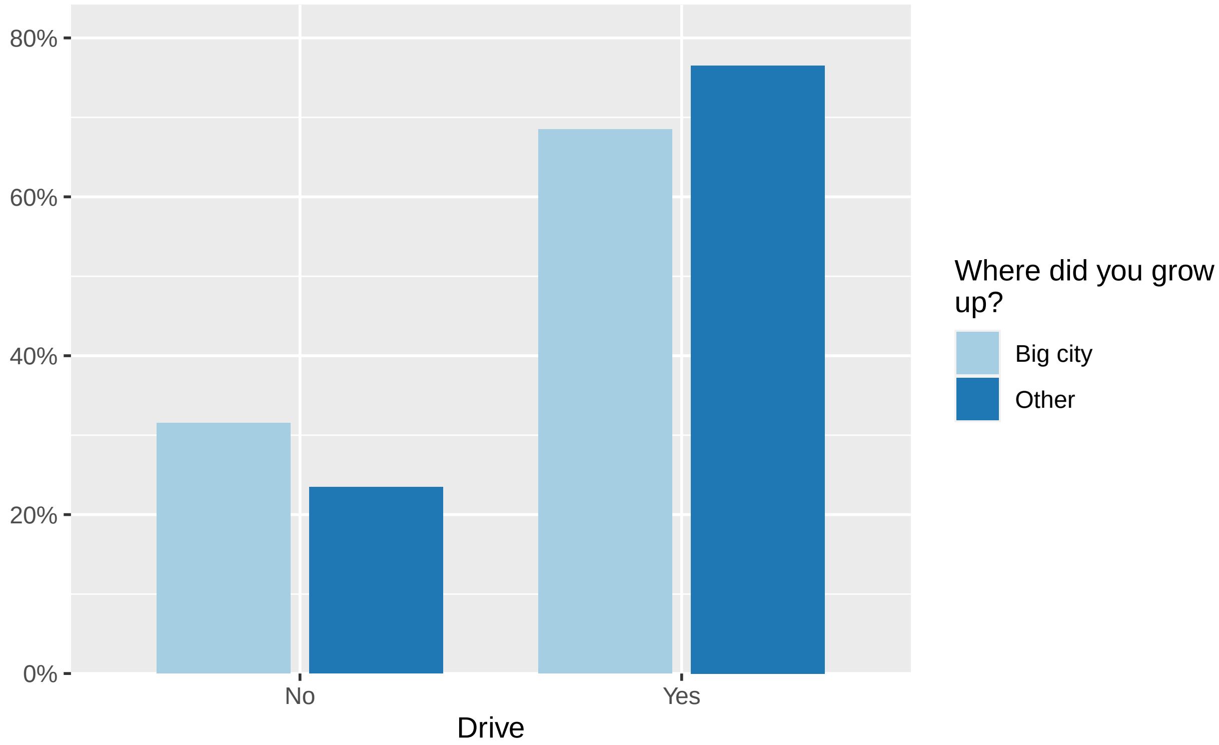
```
## Female   Male  Other
##    119     90      2
```

```
library(sjPlot)

plot_xtab(cs$Drive,cs$Gender2, show.values = FALSE,
         show.total=FALSE, axis.labels=c("No", "Yes"),
         legend.title="Gender")
```



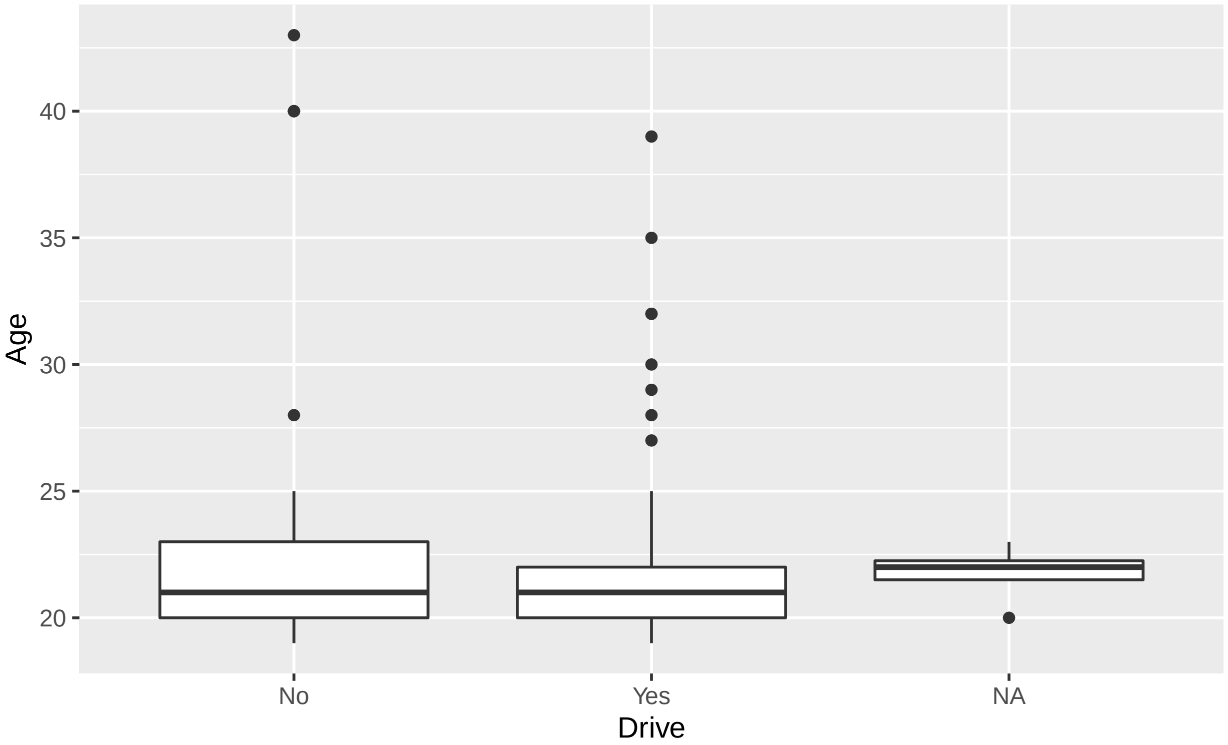
```
plot_xtab(cs$Drive,cs$GrewUpB, show.values = FALSE,
         show.total=FALSE, legend.title="Where did you grow up?")
```



There appears to be a slightly higher proportion of drivers among males than females. The proportion of drivers is also higher among students who live outside big cities.

Boxplots can be useful for exploring the relationship between the binary response and continuous predictors such as `Age`:

```
dr.plot1 <- ggplot(cs, aes(y=Age, x=Drive, group=Drive))
dr.plot1 + geom_boxplot()+ xlab("Drive")
```



Now we can fit logistic regression models to see if any of these associations are significant. Here is the model for `Age` and `GrewUpB`:

```
mod.dr <- glm(Drive ~  Age + GrewUpB, family=binomial, data=cs)
summary(mod.dr)
```

```
##
## Call:
## glm(formula = Drive ~ Age + GrewUpB, family = binomial, data = cs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8530  -1.1598   0.6876   0.8169   1.3752
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.89883    0.91665   3.162  0.00156 **
## Age           -0.09863    0.04134  -2.386  0.01704 *
## GrewUpBOther   0.49396    0.32260   1.531  0.12573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 241.69  on 206  degrees of freedom
## Residual deviance: 233.99  on 204  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 239.99
##
## Number of Fisher Scoring iterations: 4
```

We see from the output that the coefficient of age is negative (and significant), suggesting that younger respondents are slightly more likely to drive. The coefficient for `GrewUpB` is positive (but not significant), suggesting that people who grew up in small towns or rural areas are more likely to drive than those growing up in big cities.

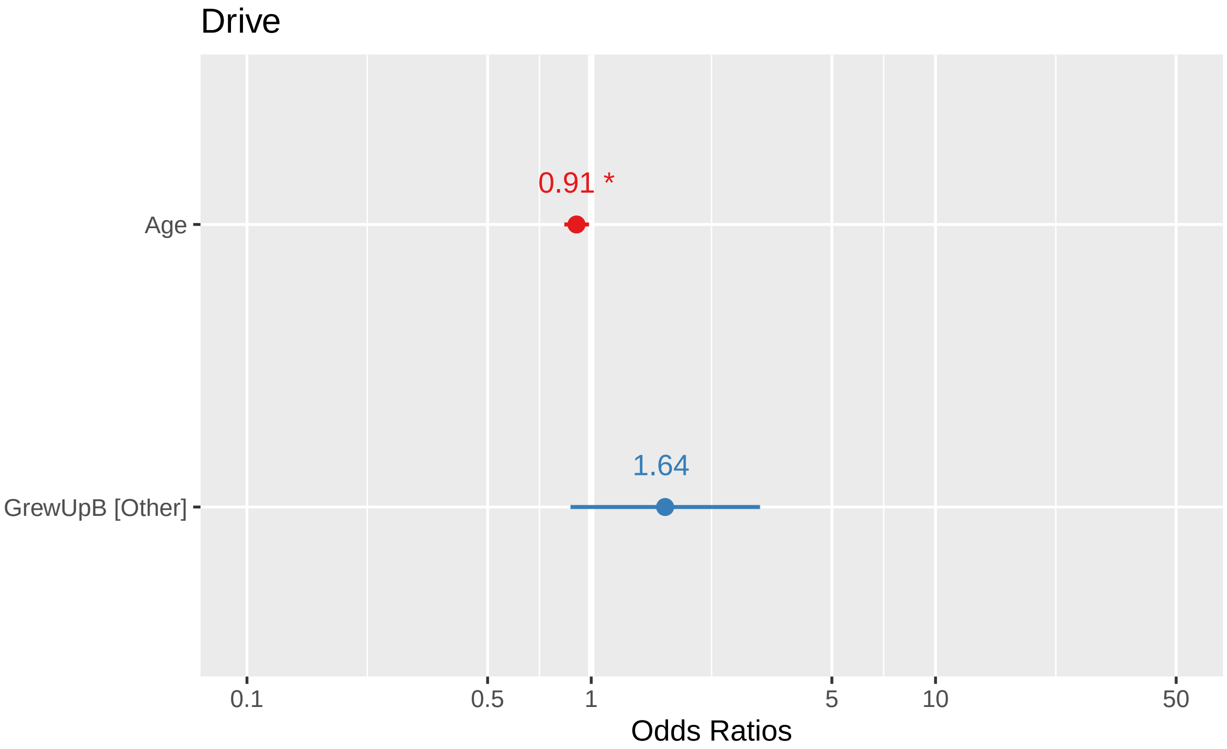
To quantify the effect of each of these predictors, we look at *odds ratios* which can be computed as $\exp(\hat{\beta})$:

```
round(exp(mod.dr$coef),2)
```

```
##  (Intercept)          Age GrewUpBOther
##        18.15         0.91         1.64
```

These are also shown in the plot below, with confidence intervals on the odds scale. The confidence interval for `GrewUpB` includes 1, while the one for `Age` does not.

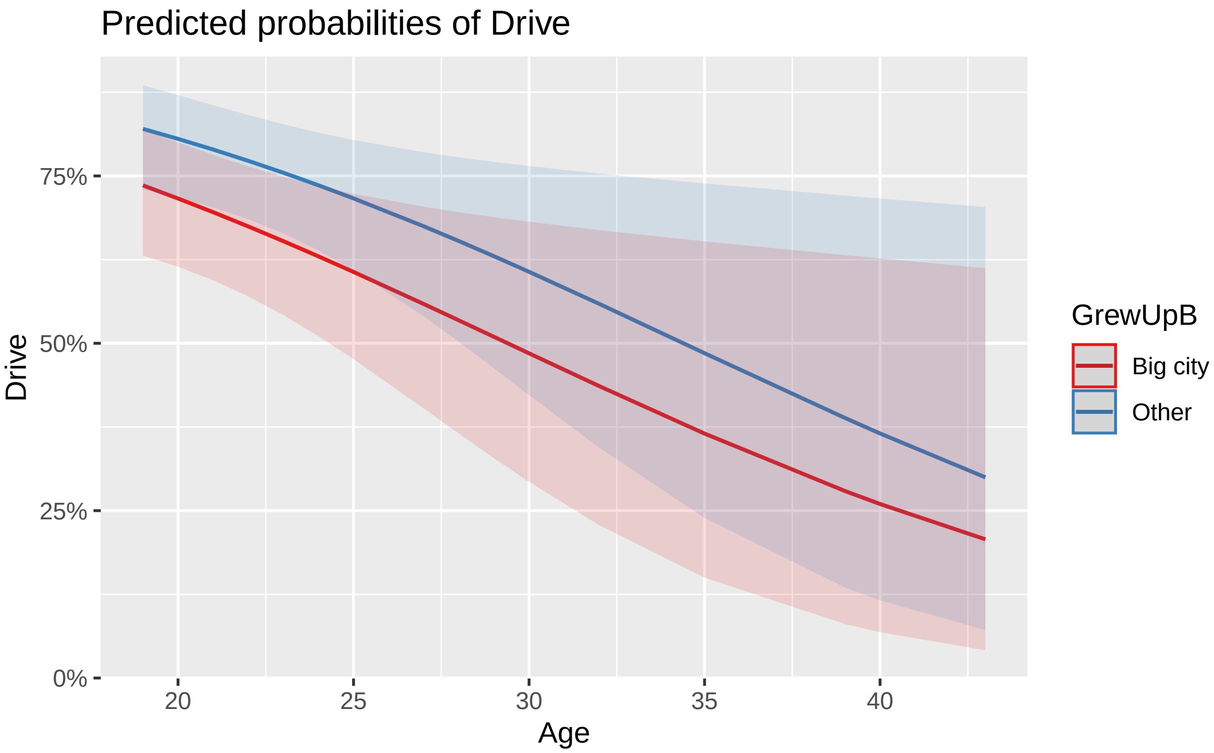
```
plot_model(mod.dr, show.values=TRUE)
```



We interpret the odds ratios as follows: For each year older, the odds of being a driver get multiplied by a factor of 0.91. The odds of driving for students from “Other” type areas are 1.64 times those of people who grew up in a big city, although this effect is not statistically significant.

We can also plot the predicted probabilities of being a driver against the student’s age by the type of place in which the student grew up.

```
plot_model(mod.dr,type="pred",terms=c("Age", "GrewUpB"))
```



On your own: Try out a model with gender as a predictor, possibly in combination with where people grew up. Is there a significant association between driving and gender?

Yanny-Laurel auditory illusion

This was done in lectures with a different dataset. Here it would be of interest to see if age is significant in predicting what people hear. First you'll have to decide how to deal with the answers that were neither "Yanny" nor "Laurel": exclude them as is done below or include them by combining them with one of the two categories?

Excluding answers other than "Yanny" and "Laurel":

```r
cs$hear3 <- factor(NA, levels=c("Laurel", "Yanny", "Other"))
cs$hear3[cs$Hear=="Laurel"] <- "Laurel"
cs$hear3[cs$Hear=="Yanny"] <-"Yanny"
cs$hear3[(cs$Hear!="Laurel"&cs$Hear!="Yanny")] <- "Other"
table(cs$hear3) # note that there is a third, empty level of this factor
```

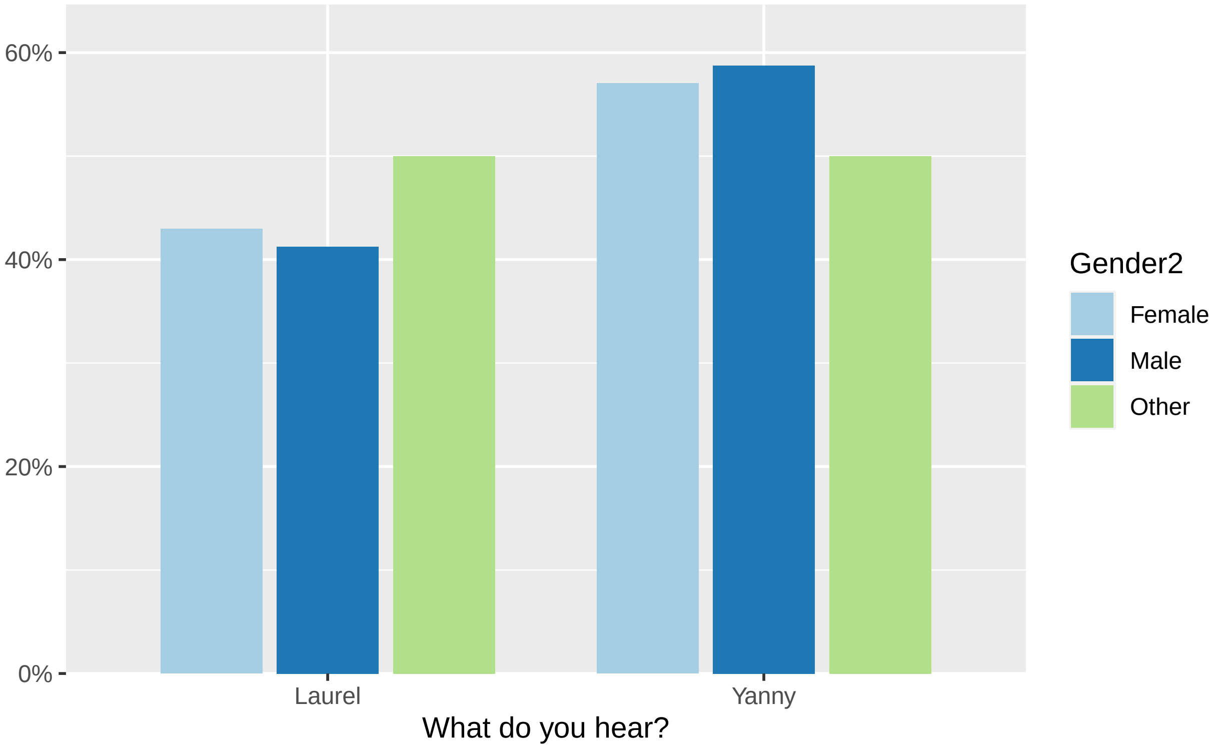
```
##
## Laurel  Yanny  Other
##     83    113     15
```

```r
yl <- cs[cs$hear3%in%c("Laurel","Yanny"),]
yl$Hear <- factor(yl$Hear) # to remove of the "Other" level
table(yl$Hear) # now empty level is gone
```

```
##
## Laurel  Yanny
##     83    113
```

The plot of the proportions against gender is shown below. A higher proportion of both males and females hear "Yanny". There were only two responses with "Other" for gender, one of which is for "Yanny" and the other for "Laurel".

```r
library(sjPlot)
plot_xtab(yl$Hear,yl$Gender2, show.values=FALSE, show.total=FALSE,
         axis.titles=c("What do you hear?"))
```



Now let us look at logistic regression models with age and gender as the explanatory variables. Here $Y_i = 1$ if the ith respondent heard "Yanny" and $Y_i = 0$ if the ith respondent heard "Laurel", with x_i being the respondent's age for $i = 1, \ldots, 194$ (excluding the two "Other" gender observations). The model we will consider is of the form

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

and we fit it in R as follows:

```r
mod.yl1 <- glm(Hear ~ Age, family=binomial, data=yl[yl$Gender2!="Other",])
summary(mod.yl1)
```

```
##
## Call:
## glm(formula = Hear ~ Age, family = binomial, data = yl[yl$Gender2 !=
##     "Other", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4134  -1.3040   0.9904   1.0229   1.8155
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.10218    0.98445   2.135   0.0327 *
## Age         -0.08225    0.04492  -1.831   0.0671 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 264.28  on 193  degrees of freedom
## Residual deviance: 260.43  on 192  degrees of freedom
## AIC: 264.43
##
## Number of Fisher Scoring iterations: 4
```

Notice that the age coefficient is negative, suggesting that older people are less likely to hear “Yanny”, but that this coefficient is not significant (*p*-value of 0.067), but could be considered “marginally significant”. Next we try a model with `Gender2`:

```
mod.yl2 <- glm(Hear ~ Gender2, family=binomial, data=yl[yl$Gender2!="Other",])
summary(mod.yl2)
```

```
##
## Call:
## glm(formula = Hear ~ Gender2, family = binomial, data = yl[yl$Gender2 !=
##     "Other", ])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.331  -1.300   1.031   1.060   1.060
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.28257    0.18919   1.494    0.135
## Gender2Male  0.07107    0.29559   0.240    0.810
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 264.28  on 193  degrees of freedom
## Residual deviance: 264.23  on 192  degrees of freedom
## AIC: 268.23
##
## Number of Fisher Scoring iterations: 4
```

The gender coefficient is positive indicating that males are more likely to hear “Yanny”, but the effect is not significant.

On your own: Repeat the analysis considering all possible models for gender and age. Are there any significant effects of either explanatory variable? Plot the estimated coefficients in the form of odds ratios and the predicted probabilities as a function of age and gender.

Also explore the following questions from the survey:

The dress/jacket

Try a similar analysis with the data for the dress or the jacket. Are any of age, gender, eye colour and/or where people grew up significant in predicting what people see? Note that you may have to create new factors with fewer levels for some of the potential predictors (e.g. eye colour).

If you would like to read a more in-depth analysis of the dress phenomenon, follow this link.

Coffee/tea

Which explanatory variables would you consider for modelling whether or not a survey participant drinks coffee (or tea)? By looking at appropriate plots and fitting appropriate logistic regression models, explore whether there are any associations between drinking coffee and age, gender or any of the other potential predictors in the data.

Astrology

Look at appropriate plots and fit logistic regression models to explore which factors (if any) are associated with an interest in astrology.