Example on the titanic data set

Gerko Vink

Supervised learning and visualization

Packages and functions used

```
library(magrittr) # pipes
library(dplyr)    # data manipulation
library(lattice)  # plotting - used for conditional plotting
library(ggplot2)  # plotting
library(ggthemes) # plotting themes
```

Titanic data

Example: titanic data

We start this lecture with a data set that logs the survival of passengers on board of the disastrous maiden voyage of the ocean liner Titanic

```
titanic <- read.csv(file = "titanic.csv", header = TRUE, stringsAsFactors = TRUE)
titanic %>% head
```

```
##   Survived Pclass                                                Name    Sex Age
## 1        0      3                             Mr. Owen Harris Braund   male  22
## 2        1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38
## 3        1      3                              Miss. Laina Heikkinen female  26
## 4        1      1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35
## 5        0      3                            Mr. William Henry Allen   male  35
## 6        0      3                                   Mr. James Moran   male  27
##   Siblings.Spouses.Aboard Parents.Children.Aboard    Fare
## 1                       1                       0  7.2500
## 2                       1                       0 71.2833
## 3                       0                       0  7.9250
## 4                       1                       0 53.1000
## 5                       0                       0  8.0500
## 6                       0                       0  8.4583
```

Inspect the data set

```
str(titanic)
```

```
## 'data.frame':    887 obs. of  8 variables:
##  $ Survived               : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass                 : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name                   : Factor w/ 887 levels "Capt. Edward Gifford Crosby",..: 602 823 172 814 733 464 7
##  $ Sex                    : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age                    : num  22 38 26 35 35 27 54 2 27 14 ...
##  $ Siblings.Spouses.Aboard: int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parents.Children.Aboard: int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare                   : num  7.25 71.28 7.92 53.1 8.05 ...
```

What sources of information

We have information on the following features.

Our outcome/dependent variable:

- Survived: yes or no

Some potential predictors:

- Sex: the passenger's gender coded as `c(male, female)`
- Pclass: the class the passenger traveled in
- Age: the passenger's age in years
- Siblings.Spouses.Aboard: if siblings or spouses were also aboard
- Parents.Children.Aboard: if the passenger's parents or children were aboard

and more.

Hypothetically

We can start investigating if there are patterns in this data that are related to the survival probability.

For example, we could hypothesize based on the crede “women and children first” that

- `Age` relates to the probability of survival in that younger passengers have a higher probability of survival
- `Sex` relates to survival in that females have a higher probability of survival

Based on socio-economic status, we could hypothesize that

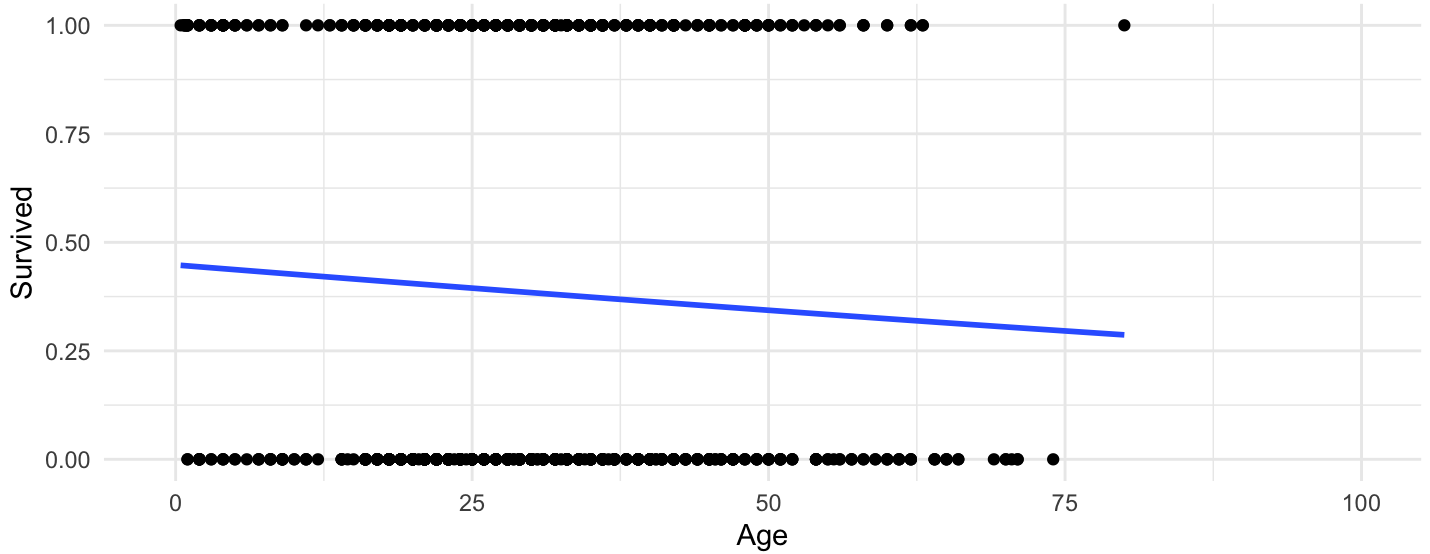
- `Pclass` relates to the probability of survival in that higher travel class leads to a higher probability of survival

And so on.

A quick investigation

Is **Age** related?

```
titanic %>% ggplot(aes(x = Age, y = Survived)) + geom_point() +
  geom_smooth(method = "glm",
    method.args = list(family = "binomial"),
    se = FALSE) + xlim(-1, 100) + theme_minimal()
```



Inspecting the data

```
titanic %$% table(Pclass, Survived)
```

```
##       Survived
## Pclass   0   1
##      1  80 136
##      2  97  87
##      3 368 119
```

It seems that the higher the class (i.e. `1 > 2 > 3`), the higher the probability of survival.

We can verify this

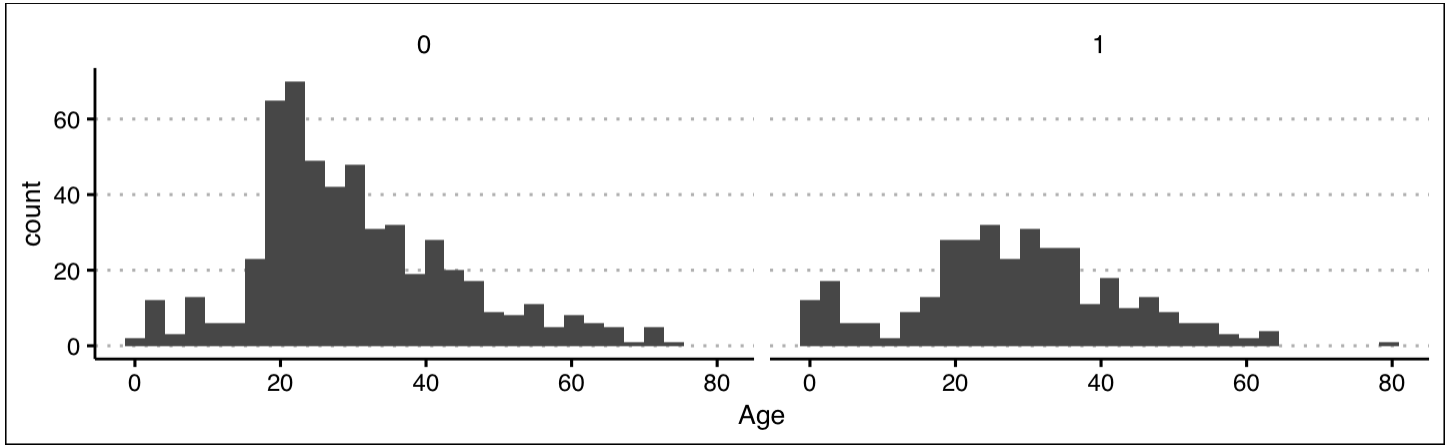
```
titanic %$% table(Pclass, Survived) %>% prop.table(margin = 1) %>% round(digits = 2)
```

```
##       Survived
## Pclass    0    1
##      1 0.37 0.63
##      2 0.53 0.47
##      3 0.76 0.24
```

A more thorough inspection

Survived ~ Age

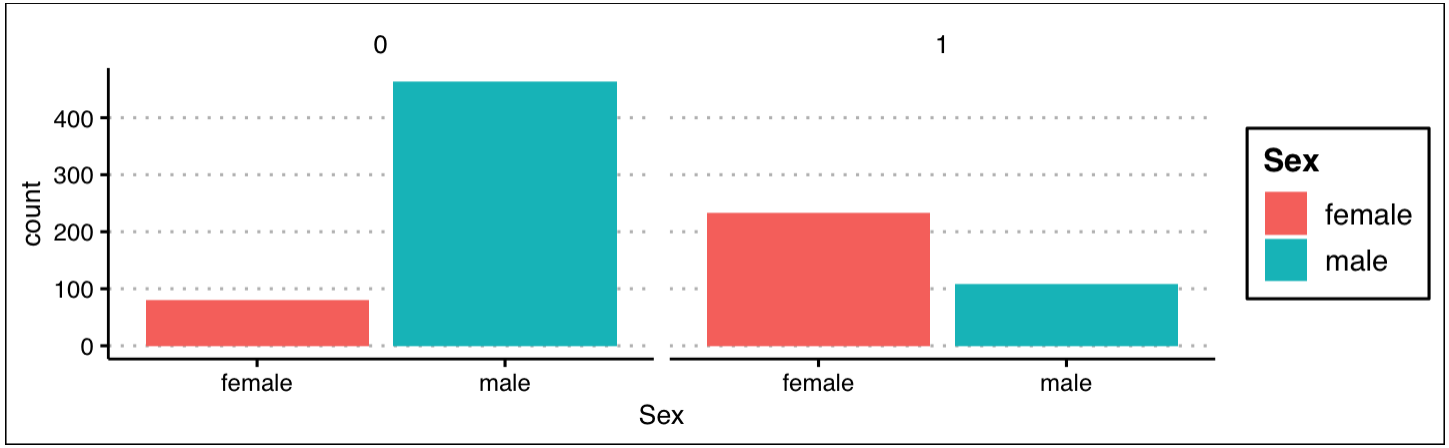
```
titanic %>%
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 30) +
  facet_wrap(~Survived) + theme_clean()
```



The distribution of `Age` for the survivors (`TRUE`) is different from the distribution of `Age` for the non-survivors (`FALSE`). Especially at the younger end there is a point mass for the survivors, which indicates that children have a higher probability of survival. However, it is not dramatically different.

Survived ~ Sex

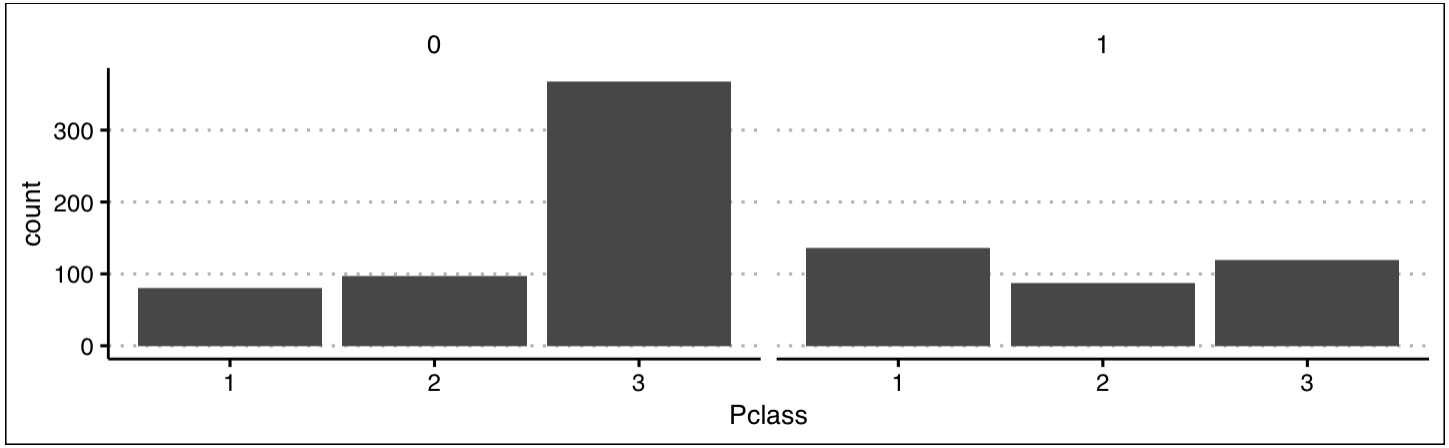
```
titanic %>%
  ggplot(aes(x = Sex)) +
  geom_bar(aes(fill = Sex)) +
  facet_wrap(~Survived) + theme_clean()
```



Wow! These distributions are very different! Females seem to have a much higher probability of survival.

Survived ~ Pclass

```
titanic %>%
  ggplot(aes(x = Pclass)) +
  geom_bar(aes(fill = Pclass)) +
  facet_wrap(~Survived) + theme_clean()
```



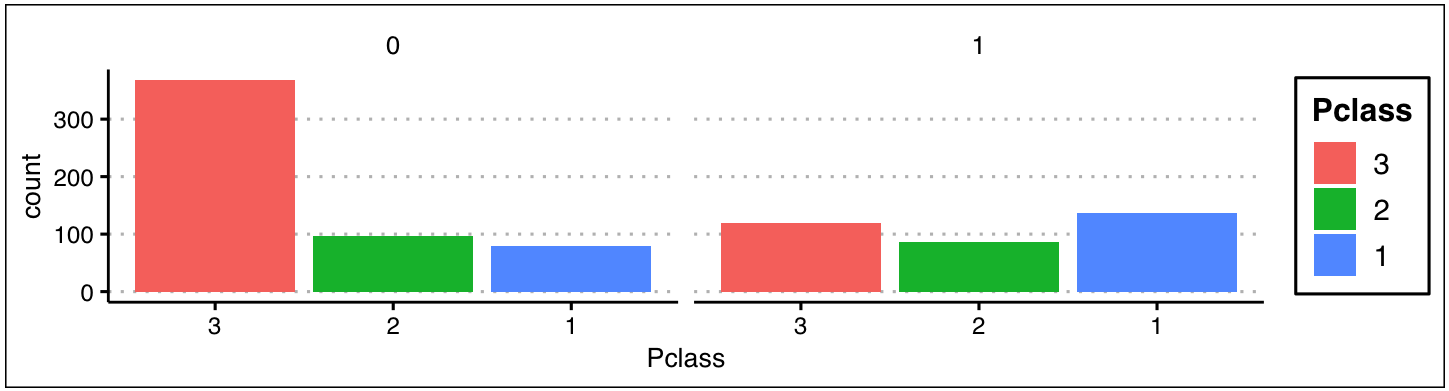
There is a very apparent difference between the distributions of the survivors and non-survivors over the classes. For example, we see that in 1st and 2nd class there are more survivors than non-survivors, while in the third class this relation is opposite.

Edit the data

```
titanic %<>%
  mutate(Pclass = factor(Pclass, levels = c(3, 2, 1), ordered = FALSE))
```

The `Pclass` column is now correctly coded as a factor. We ignore the ordering for now

```
titanic %>%
  ggplot(aes(x = Pclass)) +
  geom_bar(aes(fill = Pclass)) +
  facet_wrap(~Survived) + theme_clean()
```



Titanic with interactions

```
fit.interaction <- titanic %$% glm(Survived ~ Age * Sex * Pclass,
                                   family = binomial(link = "logit"))
fit.interaction %>% summary %>% .$coefficients
```

```
##                         Estimate Std. Error    z value    Pr(>|z|)
## (Intercept)           0.38542858 0.35158572  1.0962578 0.272965982
## Age                  -0.01742787 0.01399943 -1.2448985 0.213169059
## Sexmale              -1.24102347 0.51966698 -2.3881130 0.016935134
## Pclass2               3.66379424 1.38138966  2.6522525 0.007995671
## Pclass1               1.11218683 1.49587117  0.7435044 0.457176346
## Age:Sexmale          -0.02261191 0.02066970 -1.0939639 0.273970802
## Age:Pclass2          -0.03196246 0.03845267 -0.8312158 0.405851754
## Age:Pclass1           0.08036169 0.05283156  1.5210925 0.128236622
## Sexmale:Pclass2      -1.91119761 1.57587112 -1.2127880 0.225210878
## Sexmale:Pclass1       0.81487712 1.66024791  0.4908165 0.623556217
## Age:Sexmale:Pclass2 -0.03128163 0.04938976 -0.6333627 0.526496788
## Age:Sexmale:Pclass1 -0.08001824 0.05687997 -1.4067912 0.159489308
```

Interactions

An interaction occurs when the (causal) effect of one predictor on the outcome depends on the level of the (causal) effect of another predictor.

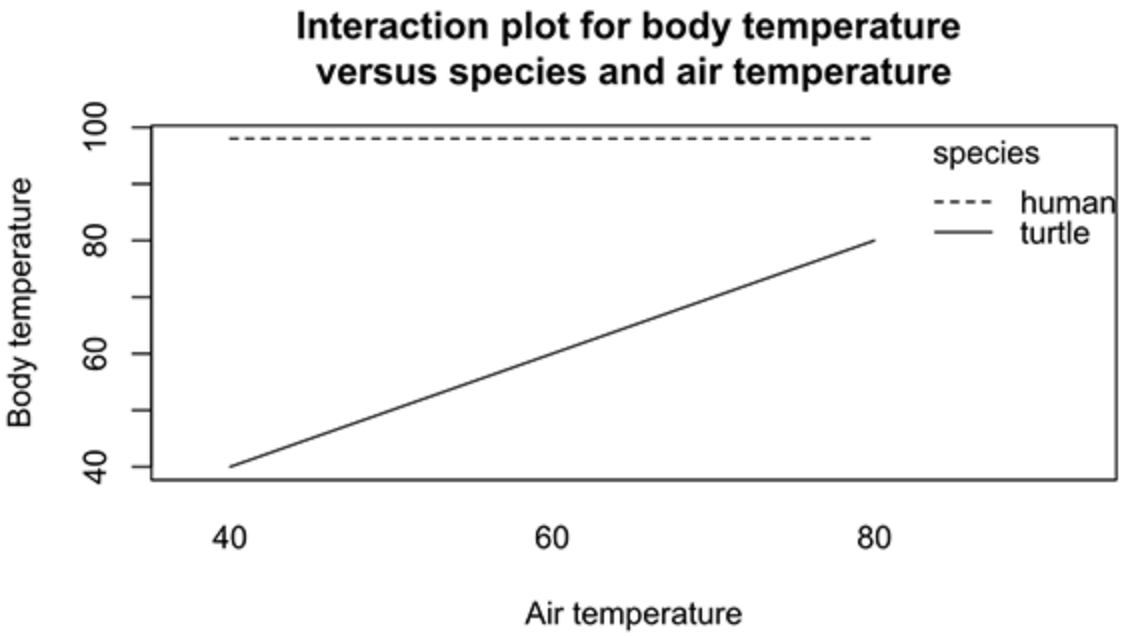


Image Source

E.g. the relation between body temperature and air temperature depends on the species.

Visualizing the effects

To illustrate, I will limit this investigation to `Age` and `Pclass` for males only.

- We can use the `predict` function to illustrate the conditional probabilities within each class

To do so, we need to create a `new` data frame that has all the combinations of predictors we need.

```
male <- data.frame(Pclass = factor(rep(c(1, 2, 3), c(80, 80, 80))),
                   Age = rep(1:80, times = 3),
                   Sex = rep("male", times = 240))
female <- data.frame(Pclass = factor(rep(c(1, 2, 3), c(80, 80, 80))),
                   Age = rep(1:80, times = 3),
                   Sex = rep("female", times = 240))
new <- rbind(female, male)
new <- cbind(new,
             predict(fit.interaction, newdata = new,
                     type = "link", se = TRUE))
```

Our **new** data set

```
head(new)
```

```
##   Pclass Age    Sex      fit   se.fit residual.scale
## 1      1   1 female 1.560549 1.407606              1
## 2      1   2 female 1.623483 1.361573              1
## 3      1   3 female 1.686417 1.315902              1
## 4      1   4 female 1.749351 1.270632              1
## 5      1   5 female 1.812285 1.225808              1
## 6      1   6 female 1.875218 1.181479              1
```

Adding the predicted probabilities

There are two simple approaches to obtain the predicted probabilities. First, we could simply ask for the predicted response:

```
new$prob <- plogis(new$fit)
head(new)
```

```
##   Pclass Age    Sex      fit   se.fit residual.scale      prob
## 1      1   1 female 1.560549 1.407606              1 0.8264322
## 2      1   2 female 1.623483 1.361573              1 0.8352749
## 3      1   3 female 1.686417 1.315902              1 0.8437524
## 4      1   4 female 1.749351 1.270632              1 0.8518709
## 5      1   5 female 1.812285 1.225808              1 0.8596378
## 6      1   6 female 1.875218 1.181479              1 0.8670609
```

Adding confidence intervals

```
new %<>%
  mutate(lower = plogis(fit - 1.96 * se.fit),
         upper = plogis(fit + 1.96 * se.fit))

head(new)
```

```
##   Pclass Age    Sex      fit   se.fit residual.scale      prob     lower
## 1      1   1 female 1.560549 1.407606              1 0.8264322 0.2317674
## 2      1   2 female 1.623483 1.361573              1 0.8352749 0.2601478
## 3      1   3 female 1.686417 1.315902              1 0.8437524 0.2905423
## 4      1   4 female 1.749351 1.270632              1 0.8518709 0.3227661
## 5      1   5 female 1.812285 1.225808              1 0.8596378 0.3565664
## 6      1   6 female 1.875218 1.181479              1 0.8670609 0.3916264
##       upper
## 1 0.9868676
## 2 0.9865092
## 3 0.9861508
## 4 0.9857941
## 5 0.9854408
## 6 0.9850932
```

What do we have?

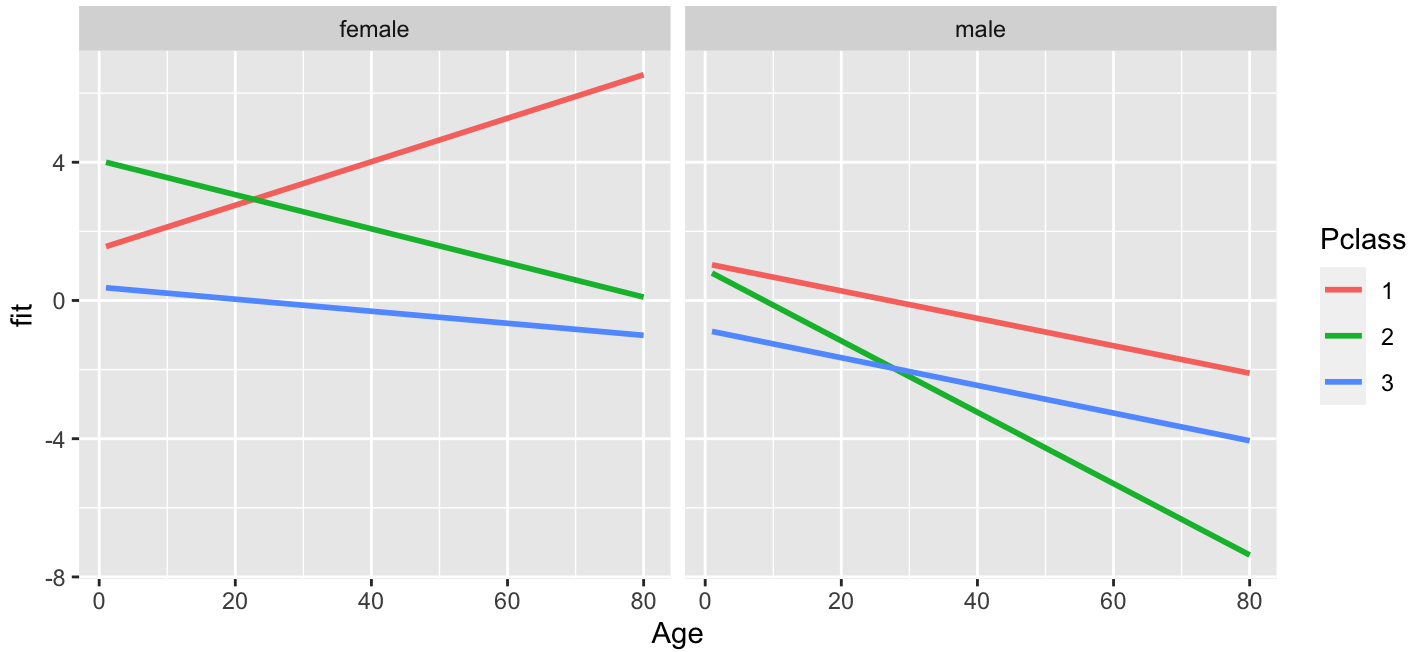
A data frame with simulated `Pclass` and `Age` for males.

```
new %>% summary()
```

```
##  Pclass       Age            Sex                 fit               se.fit       
##  1:160   Min.   : 1.00   Length:480         Min.   :-7.36571   Min.   :0.1588  
##  2:160   1st Qu.:20.75   Class :character   1st Qu.:-1.78710   1st Qu.:0.3228  
##  3:160   Median :40.50   Mode  :character   Median :-0.25069   Median :0.5526  
##          Mean   :40.50                      Mean   :-0.08741   Mean   :0.6962  
##          3rd Qu.:60.25                      3rd Qu.: 1.81588   3rd Qu.:0.8838  
##          Max.   :80.00                      Max.   : 6.53232   Max.   :2.8111  
##  residual.scale      prob              lower               upper          
##  Min.   :1      Min.   :0.0006322   Min.   :0.0000271   Min.   :0.01454  
##  1st Qu.:1      1st Qu.:0.1434293   1st Qu.:0.0709389   1st Qu.:0.25291  
##  Median :1      Median :0.4376551   Median :0.2512664   Median :0.60256  
##  Mean   :1      Mean   :0.4769910   Mean   :0.3398537   Mean   :0.58495  
##  3rd Qu.:1      3rd Qu.:0.8600691   3rd Qu.:0.5426823   3rd Qu.:0.96181  
##  Max.   :1      Max.   :0.9985465   Max.   :0.9033353   Max.   :0.99999
```

Visualizing the effects: link

```
new %>%
  ggplot(aes(x = Age, y = fit)) +
  geom_line(aes(colour = Pclass), lwd = 1) +
  facet_wrap(~ Sex)
```



Visualizing the effects: probabilities

```
new %>%
  ggplot(aes(x = Age, y = prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = Pclass), alpha = .2) +
  geom_line(aes(colour = Pclass), lwd = 1) + ylab("Probability of Survival") +
  facet_wrap(~ Sex)
```