w2ex3

Paul Hosek

2023-03-03

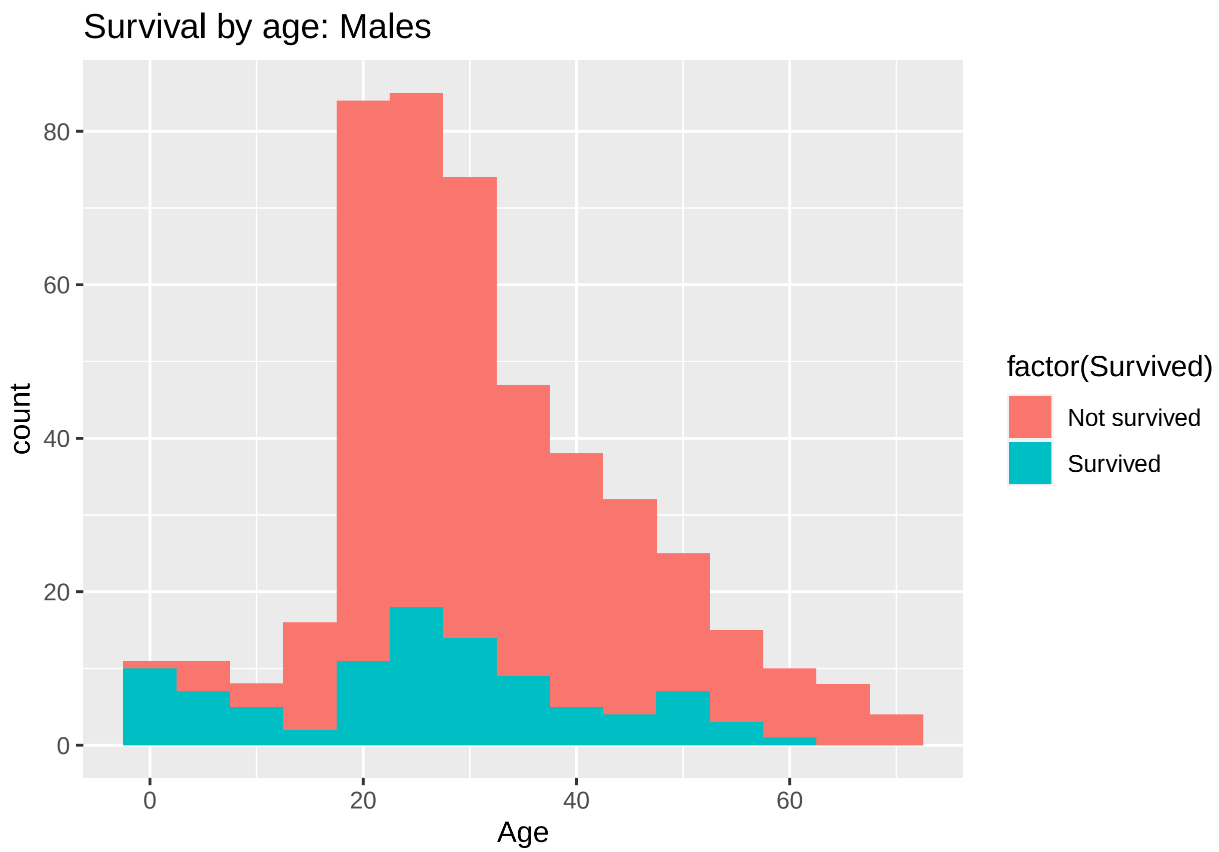
- models will only work with !nan ages

#3 ## a) - multiple summaries of data - fit log regression,, w/o interactions -> survival & predictor Pclass, age, sex

```
data_titanic <- read.table("titanic.txt", header=TRUE)
data_titanic$PClass <- as.factor(data_titanic$PClass)
data_titanic$Sex <- as.factor(data_titanic$Sex)
data_titanic$Survived <- as.factor(data_titanic$Survived)
```

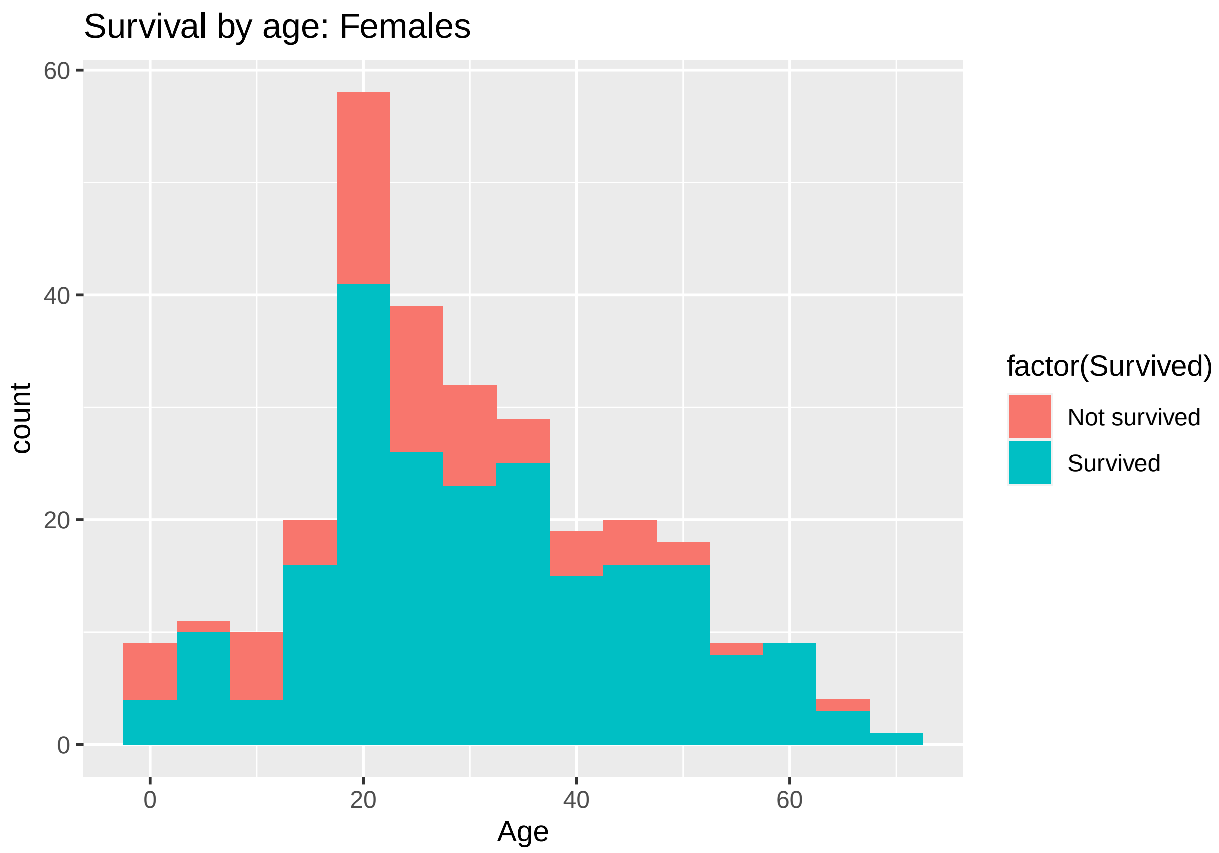
```
par(mfrow=c(1,3))

ggplot(subset(data_titanic, Sex == "male"),
       aes(x = Age, fill = factor(Survived))) +geom_histogram(binwidth = 5)+scale_fill_discrete(labels
```

```
## Warning: Removed 383 rows containing non-finite values (`stat_bin()`).
```

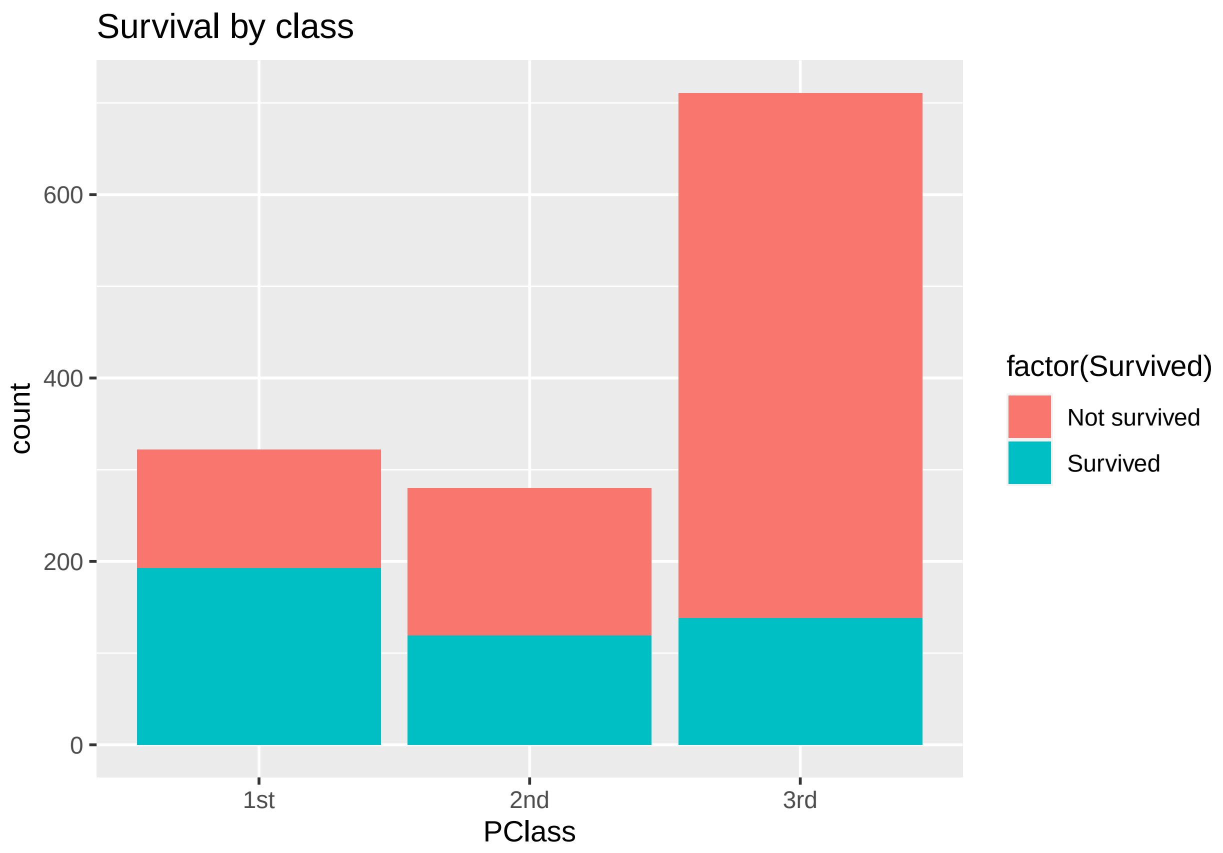


```
ggplot(subset(data_titanic, Sex == "female"), aes(x = Age, fill = factor(Survived))) +
  geom_histogram(binwidth = 5)+scale_fill_discrete(labels = c("Not survived", "Survived"))+labs(title =
```

```
## Warning: Removed 174 rows containing non-finite values (`stat_bin()`).
```



```
ggplot(data_titanic, aes(x = PClass, fill = factor(Survived))) +
  geom_bar()+scale_fill_discrete(labels = c("Not survived", "Survived"))+labs(title = "Survival by clas
```



```
model_log1 <- glm(Survived ~ PClass + Age + Sex, data = data_titanic, family = binomial())
summary(model_log1)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age + Sex, family = binomial(),
##     data = data_titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7226  -0.7065  -0.3917   0.6495   2.5289
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.759662   0.397567   9.457  < 2e-16 ***
## PClass2nd   -1.291962   0.260076  -4.968 6.78e-07 ***
## PClass3rd   -2.521419   0.276657  -9.114  < 2e-16 ***
## Age         -0.039177   0.007616  -5.144 2.69e-07 ***
## Sexmale     -2.631357   0.201505 -13.058  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  695.14  on 751  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 705.14
##
## Number of Fisher Scoring iterations: 5
```

Excluding interaction effects, we find that being a female or a first class passengers or young increases your odds of survival. However, we cannot know how a combination of these will impact the odds. From the main effects we can conclude: Males are 13.89 more likely to die compared to females. 2nd-class passengers are 3.64 and third-class passengers are 12.45 as likely to die than passengers in other classes (calculated as 1/exp(coefficient_of_interest)). Further, for each year a person is older, odds decrease by a factor of 0.96: younger passengers are more likely to survive (calculated as exp(age)). All these main effects are statistically significantly associated with survival. ## b)

```
model_log2 <- glm(Survived ~ PClass + Age + Sex + PClass:Age + Age:Sex, data = data_titanic, family = bi
summary(model_log2)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age + Sex + PClass:Age + Age:Sex,
##     family = binomial, data = data_titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6858  -0.6459  -0.3392   0.6751   2.7271
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.02992    0.65977   3.077  0.00209 **
## PClass2nd     -0.21153    0.71014  -0.298  0.76580
## PClass3rd     -2.08114    0.66578  -3.126  0.00177 **
## Age            0.02459    0.01975   1.245  0.21310
## Sexmale       -0.38894    0.48027  -0.810  0.41804
## PClass2nd:Age -0.04506    0.02195  -2.053  0.04012 *
## PClass3rd:Age -0.01481    0.02113  -0.701  0.48337
## Age:Sexmale   -0.08209    0.01707  -4.809 1.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  662.45  on 748  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 678.45
##
## Number of Fisher Scoring iterations: 5
```

```
all_comb_55 <- expand.grid(PClass = levels(data_titanic$PClass), Sex = levels(data_titanic$Sex), Age = 5
all_comb_55$Survival_Probability <- predict(model_log2, all_comb_55, type = "response") # response = pr
kable(all_comb_55, format = "latex", caption = "Survival probability for 55 year olds.")
```

We observe that being female has the largest influence on survival. Independent of gender, more expensive classes have larger survival probability. We observe that females in the first class have a 0.97

Table 1: Survival probability for 55 year olds.

PClass	Sex	Age	Survival_Probability
1st	female	55	0.9671529
2nd	female	55	0.6665224
3rd	female	55	0.6193971
1st	male	55	0.1792333
2nd	male	55	0.0146069
3rd	male	55	0.0119258