els in R, Part 2

Scenario

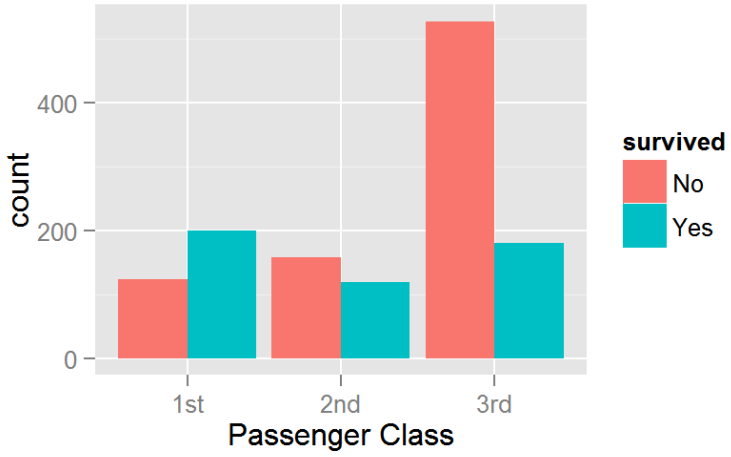
- Imagine you are about to board the Titanic
- Who would you rather be - Jack or Rose?
- Who has the highest probability of surviving?

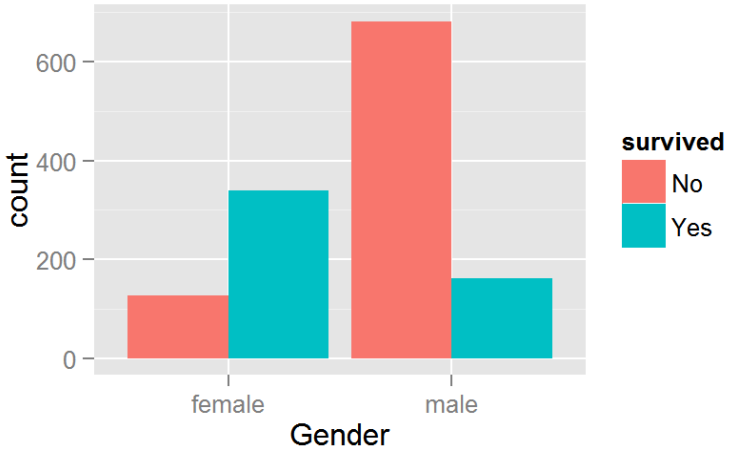
Titanic Survival Data

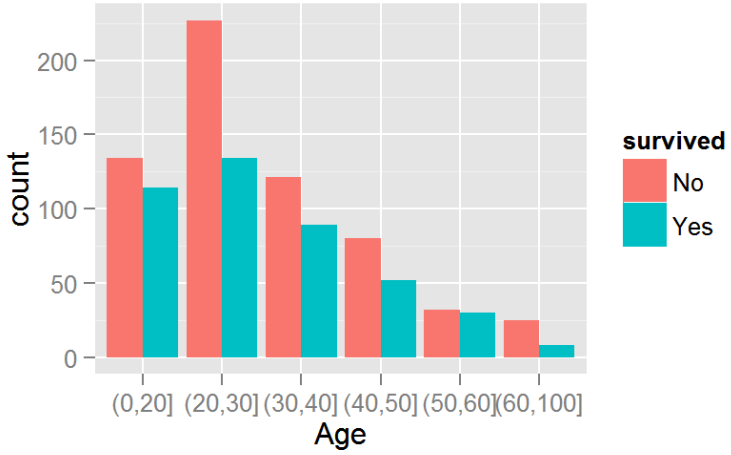
```
## load data
load('data/titanic2.rda')
head(titanic)
```

```
##   pclass survived                             name    sex     age sibsp
## 1    1st      Yes    Allen, Miss. Elisabeth Walton female 29.0000     0
## 2    1st      Yes   Allison, Master. Hudson Trevor   male  0.9167     1
## 3    1st       No     Allison, Miss. Helen Loraine female  2.0000     1
## 4    1st       No Allison, Mr. Hudson Joshua Crei   male 30.0000     1
## 5    1st       No Allison, Mrs. Hudson J C (Bessi female 25.0000     1
## 6    1st      Yes              Anderson, Mr. Harry   male 48.0000     0
##   parch ticket     fare   cabin    embarked boat body
## 1     0  24160 211.3375      B5 Southampton    2   NA
## 2     2 113781 151.5500 C22 C26 Southampton   11   NA
## 3     2 113781 151.5500 C22 C26 Southampton        NA
## 4     2 113781 151.5500 C22 C26 Southampton       135
## 5     2 113781 151.5500 C22 C26 Southampton        NA
## 6     0  19952  26.5500     E12 Southampton    3   NA
##                         home.dest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                     New York, NY
```

Survival Counts







Logit Model

- A model for predicting the outcome of binary event
- Model describes probability of outcome as a function of X variables
- The Logit model is

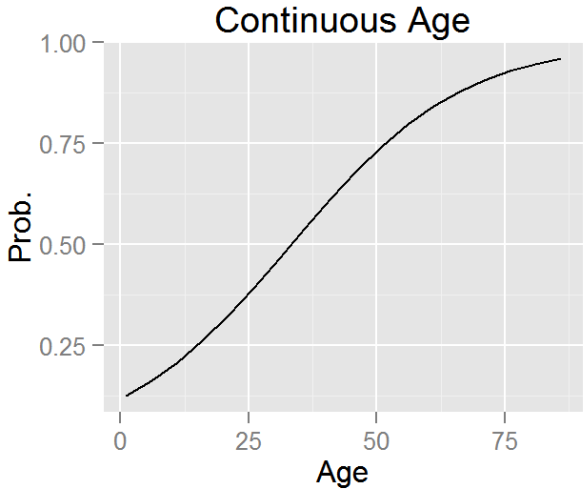
$$\Pr(\mathrm{Y}=1|\mathrm{X}) = \frac{\exp\{\beta_1 X_1 + \beta_2 X_2 + \ldots\}}{1 + \exp\{\beta_1 X_1 + \beta_2 X_2 + \ldots\}}$$

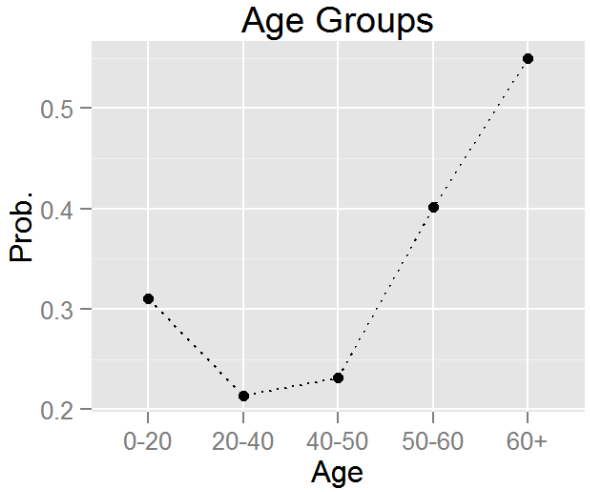
- If we know β we can predict probabilities for each X
- In the Titanic example we can let X consist of gender, age and class of travel

Example

$$\Pr(\mathrm{Y}=1|\mathrm{X}) = \frac{\exp\{-2+0.06*age\}}{1+\exp\{-2+0.06*age\}}$$

$$\Pr(\mathrm{Y}=1|\mathrm{X}) = \frac{\exp\{-0.8-0.5D_{0-20}-0.4D_{20-40}\ldots\}}{1+\exp\{-0.8-0.5D_{0-20}-D_{20-40}\ldots\}}$$





Logit Model in R

```
## define age groups
titanic.est <- titanic %>%
  filter(!age=='NA') %>%
  mutate(age.f=cut(age,breaks=c(0,20,30,40,50,60,100)))

## define logit model
logit.titanic.A <- glm(survived=='Yes'~pclass+sex+age.f,
                       data=titanic.est,
                       family=binomial(link="logit"))
```

- Here we defined the dependent variable as a logical
- We are modelling the probability that the logical is TRUE (i.e., person survived)
- R will calibrate the β's to give the best fit to the data

What do the β's look like?

```
summary(logit.titanic.A)$coefficients
```

```
##                 Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)    2.9320018  0.2751003  10.657938 1.601075e-26
## pclass2nd     -1.1920902  0.2254817  -5.286860 1.244338e-07
## pclass3rd     -2.1522367  0.2228720  -9.656828 4.598813e-22
## sexmale       -2.4875254  0.1652664 -15.051610 3.369334e-51
## age.f(20,30]  -0.4999483  0.2120267  -2.357950 1.837618e-02
## age.f(30,40]  -0.4944054  0.2467401  -2.003750 4.509687e-02
## age.f(40,50]  -1.0336425  0.2890050  -3.576556 3.481513e-04
## age.f(50,60]  -1.0883060  0.3854590  -2.823402 4.751689e-03
## age.f(60,100] -1.9086657  0.5393778  -3.538644 4.021885e-04
```

- What does this mean?
- Start with the baseline passenger: 1st class, female, 20 years or younger. The predicted survival probability for this group is

$$Pr(Surv|base) = \frac{\exp(2.932)}{1.0 + \exp(2.932)} \approx 0.95$$

Interpreting the β's

- Suppose the same person (female, 20 years or younger) travelled on 2nd class?

$$Pr(Surv|female, age20, 2ndclass) = \frac{\exp(2.932 - 1.192)}{1.0 + \exp(2.932 - 1.192)} \approx 0.85$$

- 3rd Class:

$$Pr(Surv|female, age20, 3rdclass) = \frac{\exp(2.932 - 2.15)}{1.0 + \exp(2.932 - 2.15)} \approx 0.69$$

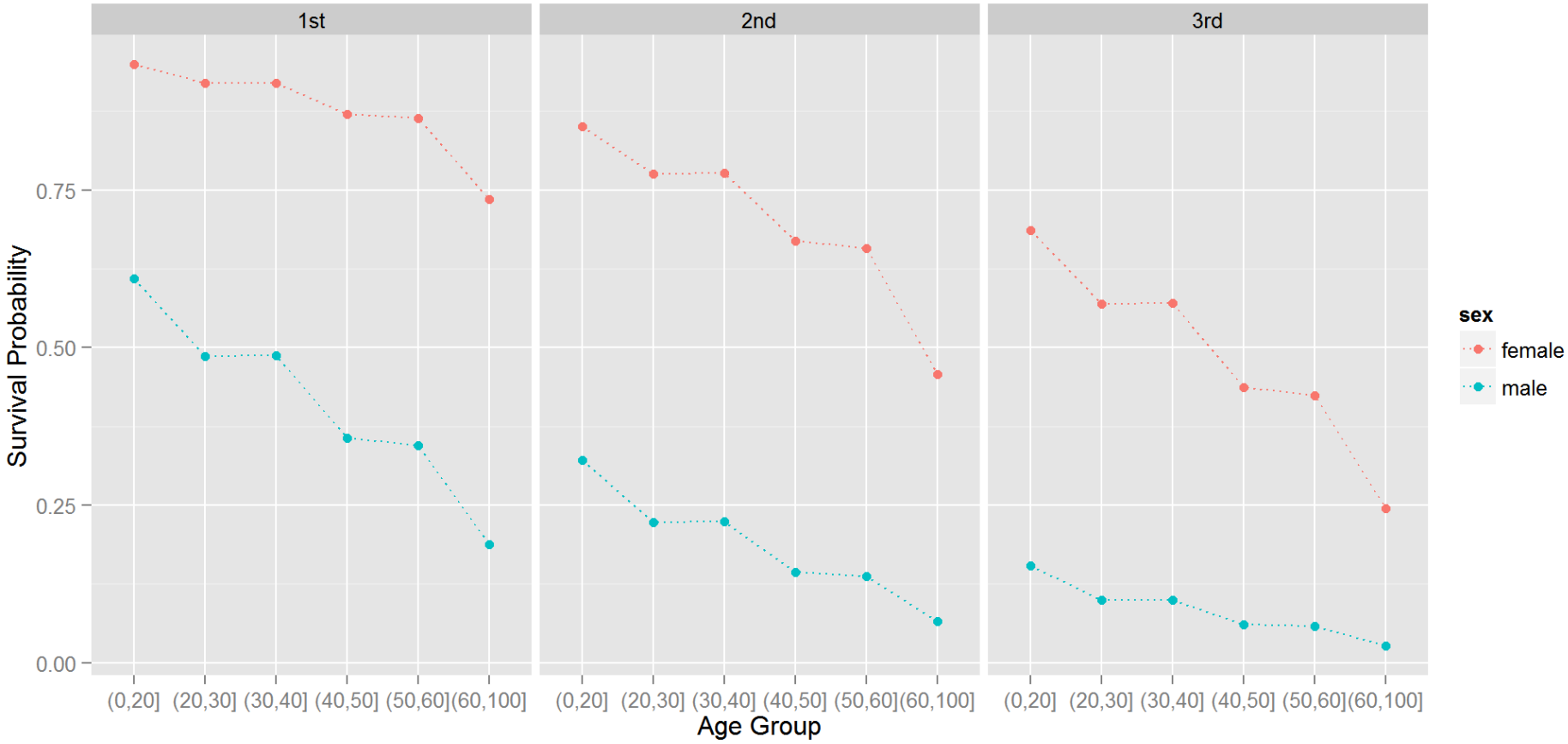
Predicting Survival for All Groups

```
## create prediction array
pred.df <- expand.grid(pclass=levels(titanic.est$pclass),
                       sex=levels(titanic.est$sex),
                       age.f=levels(titanic.est$age.f))

## predictions
pred.df$Prob <- predict(logit.titanic.A, newdata = pred.df, type = "response")
```

- The array pred.df contains the data for the passenger groups for which we want to form predictions
- The **predict** command will make the predictions for these group using the logit model
- Visualize the predictions:

```
ggplot(data=pred.df,aes(x=age.f,y=Prob,group=sex,color=sex)) + geom_line(linetype='dotted') +
  geom_point() + ylab('Survival Probability') + xlab('Age Group') + facet_wrap(~pclass)
```

Predictions



Plot of β estimates

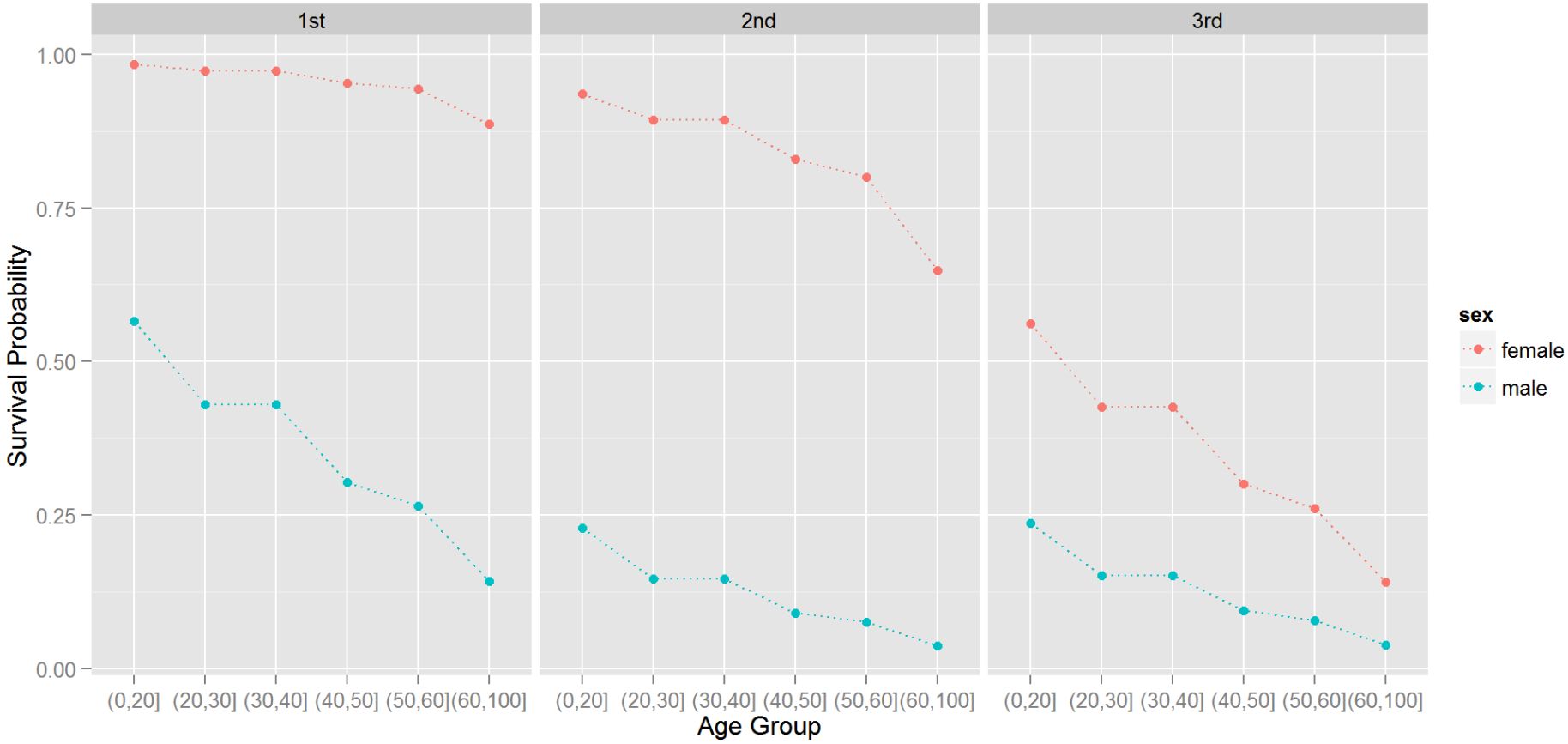
```
sum.c <- summary(logit.titanic.A)$coefficients

estimates.logit.A <- data.frame(
  Parameter=rownames(sum.c),
  Estimate=sum.c[,'Estimate'],
  confint.default(logit.titanic.A))

ggplot(data=filter(estimates.logit.A,
                   !Parameter=='(Intercept)'),
  aes(x=Parameter,y=Estimate,
      ymin=X2.5..,ymax=X97.5..)) +
  geom_pointrange() +
  coord_flip() +
  xlab(' ') +
  geom_hline(xintercept=0)
```

Adding an Interaction

- Question: Is the effect passenger class different for males and females?

```
logit.titanic.B <- glm(survived=='Yes'~pclass*sex+age.f,
                       data=titanic.est,
                       family=binomial(link="logit"))
```

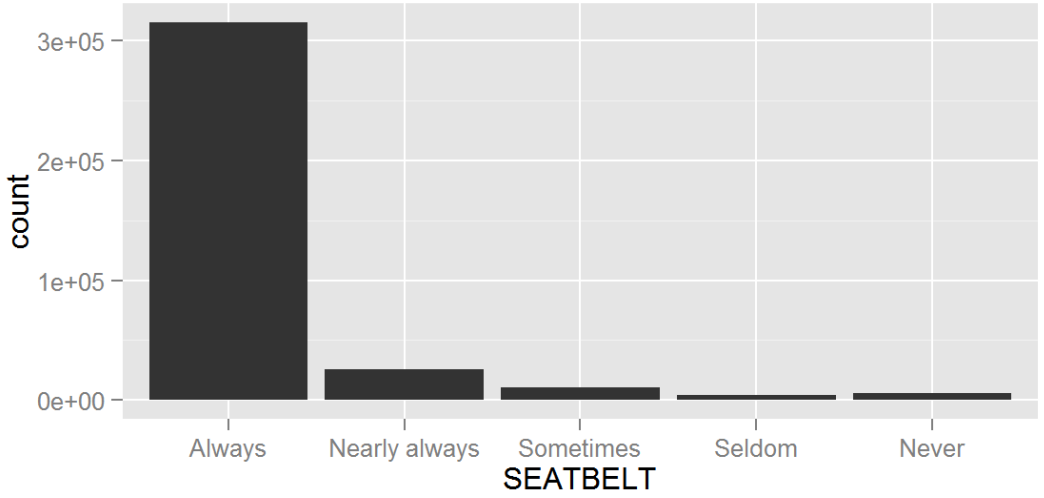
Result


1st
2nd
3rd
1.00
0.75
0.50
0.25
0.00
Survival Probability
(0,20]
(20,30]
(30,40]
(40,50]
(50,60]
(60,100]
Age Group
sex
female
male

Case Study: Who doesn't wear seat belts?

Objectives

- You have been tasked with designing a marketing communications campaign to promote seat bealt usage
- Who should this campaign target?
- Data: 2011 wave of the Behavioral Risk Factor Surveillance System (**brfs11_sub.rda**)
- Annual survey conducted by the CDC
- 361,836 respondents
- Key question: "How often do you use seat belts when you drive or ride in a car?"



Data Set-up

```
load('data/brfss11_sub.rda')
names(brfss11)
```

```
## [1] "SEATBELT" "EDU"      "INCOME"   "Red_Blue" "AGE"      "State"
```

- EDU=Highest level of education of respondent
- INCOME=Annual household income of respondent
- AGE=Age group of respondent
- Red_Blue=Political orientiation of respondent's home state
- State=Name of respondent's home state

Click Analytics

- The file **clicks.rda** contains 973,683 instances of online users being exposed to an ad.
- The data is colleced over a 4 hour period.
- This data is heavily masked for proprietary reasons
- Variables are

1. click = Did the user click on the ad? (1=yes,0=no)
2. hour = Index for hour
3. banner_pos = Position of ad on page
4. site_id = web site id
5. site_category = web site category
6. app_category = Application category
7. device_os = Operation sytem of device 8-11. C18,21,C24 = Masked variables

Can You Predict Clicks?

You are web developer interested in making advertising dollars. What type of web site ("category") should you develop?