Advanced Applied Statistics Homework 7 Solutions

2022-11-22

1. Logistic Regression

Access the Washington Post dataset on murder crimes (available on Canvas), and fit a logistic regression model, using as response variable whether or not a case is closed. (Be careful, when I read in the data, everything was read in as a character vector. I had to turn variables like age into a numeric vector by using `as.numeric`. Also, you'll have to create the response variable from the existing variable called `disposition`, which has the three categories "Open/No arrest", "Closed without arrest", "Closed by arrest". Note: There are relatively few cases that are "Closed without arrest". If you want, instead of merging the two "Closed" categories, you can also filter out the ones that are "Closed without arrest", and just consider the binary response "Open/No arrest" versus "Closed by arrest".)

a. Make some preliminary plots, showing the distribution of the response variable, the distribution of the variable `sex`, and the distribution of the variable `age`.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
crime <- read_csv("/Users/joshuaingram/Main/Projects/masters_coursework/teaching_assistant/advanced_appl
```

```
## Rows: 52179 Columns: 12
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): uid, victim_last, victim_first, victim_race, victim_age, victim_sex...
## dbl (3): reported_date, lat, lon
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
crime
```

```
## # A tibble: 52,179 x 12
##    uid        reported_date victim_last  victim_first victim_race victim_age
##    <chr>              <dbl> <chr>        <chr>        <chr>       <chr>
##  1 Alb-000001      20100504 GARCIA       JUAN         Hispanic    78
##  2 Alb-000002      20100216 MONTOYA      CAMERON      Hispanic    17
##  3 Alb-000003      20100601 SATTERFIELD  VIVIANA      White       15
##  4 Alb-000004      20100101 MENDIOLA     CARLOS       Hispanic    32
##  5 Alb-000005      20100102 MULA         VIVIAN       White       72
##  6 Alb-000006      20100126 BOOK         GERALDINE    White       91
```

```
##  7 Alb-000007      20100127 MALDONADO    DAVID        Hispanic    52
##  8 Alb-000008      20100127 MALDONADO    CONNIE       Hispanic    52
##  9 Alb-000009      20100130 MARTIN-LEYVA GUSTAVO      White       56
## 10 Alb-000010      20100210 HERRERA      ISRAEL       Hispanic    43
## # ... with 52,169 more rows, and 6 more variables: victim_sex <chr>,
## #   city <chr>, state <chr>, lat <dbl>, lon <dbl>, disposition <chr>
```

```
crime %>% count(disposition)
```

```
## # A tibble: 3 x 2
##   disposition               n
##   <chr>                 <int>
## 1 Closed by arrest      25674
## 2 Closed without arrest  2922
## 3 Open/No arrest        23583
```

```
crime1 <- crime %>%
  mutate(disposition = factor(disposition,
                              levels=c("Open/No arrest", "Closed without arrest", "Closed by arrest"), 
         disposition.binary = fct_collapse(disposition, "Open" = "Open/No arrest", "Closed" = c("Closed
         age = as.numeric(victim_age),
         race = factor(victim_race),
         sex = factor(victim_sex)
  ) %>%
  select("disposition.binary", "age", "race", "sex") %>%
  drop_na()
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
crime1
```

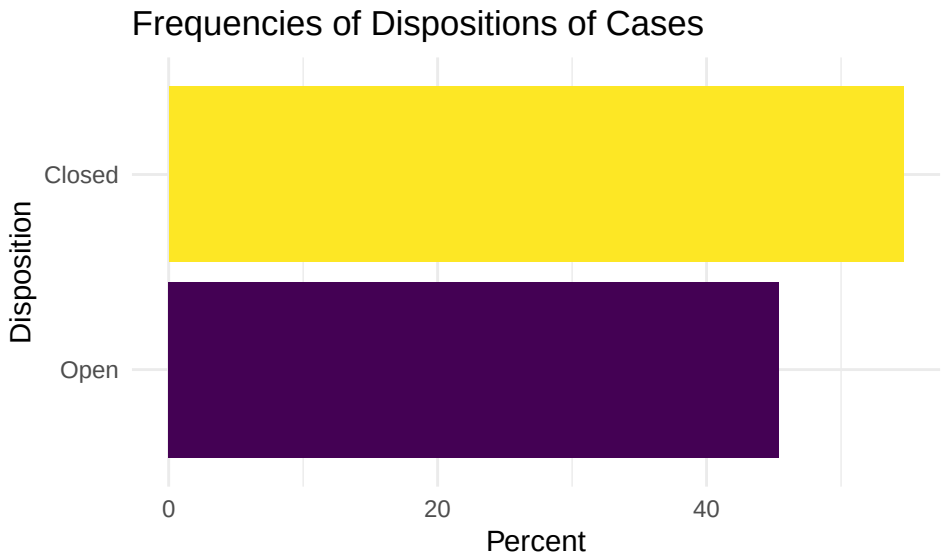
```
## # A tibble: 49,180 x 4
##    disposition.binary   age race     sex
##    <ord>              <dbl> <fct>    <fct>
##  1 Closed                78 Hispanic Male
##  2 Closed                17 Hispanic Male
##  3 Closed                15 White    Female
##  4 Closed                32 Hispanic Male
##  5 Closed                72 White    Female
##  6 Open                  91 White    Female
##  7 Closed                52 Hispanic Male
##  8 Closed                52 Hispanic Female
##  9 Open                  56 White    Male
## 10 Open                  43 Hispanic Male
## # ... with 49,170 more rows
```

```
crime1 %>% count(disposition.binary)
```

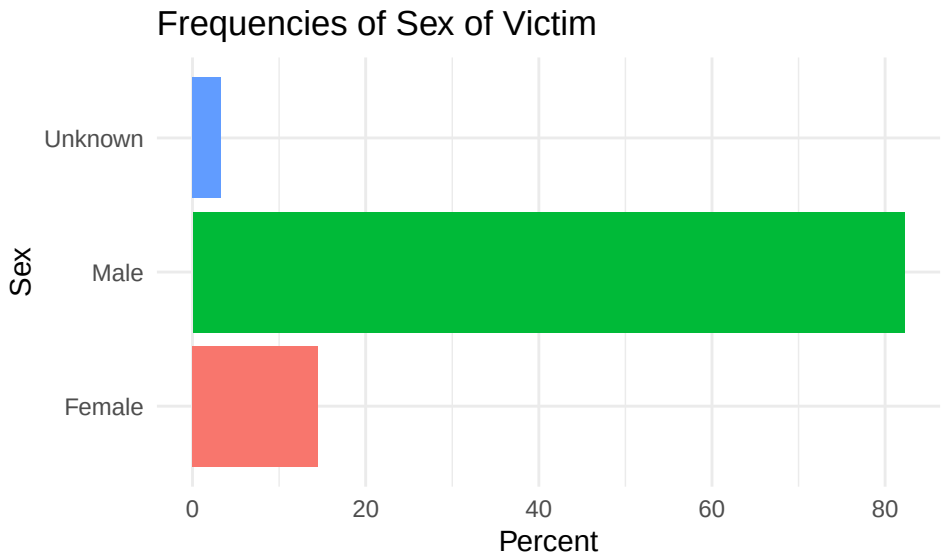
```
## # A tibble: 2 x 2
##   disposition.binary     n
##   <ord>              <int>
## 1 Open               22311
## 2 Closed             26869
```

```
crime1 %>% count(disposition.binary) %>%
  ggplot(aes(x=disposition.binary, y=100*n/sum(n), fill=disposition.binary)) +
    geom_bar(stat="identity", show.legend=FALSE) +
```

```
    labs(x="Disposition",
         y="Percent",
         title="Frequencies of Dispositions of Cases") +
    theme_minimal() +
  coord_flip()
```




```
crime1 %>% count(sex) %>%
  ggplot(aes(x=sex, y=100*n/sum(n), fill=sex)) +
    geom_bar(stat="identity", show.legend=FALSE) +
    labs(x="Sex",
         y="Percent",
         title="Frequencies of Sex of Victim") +
    theme_minimal() +
  coord_flip()
```

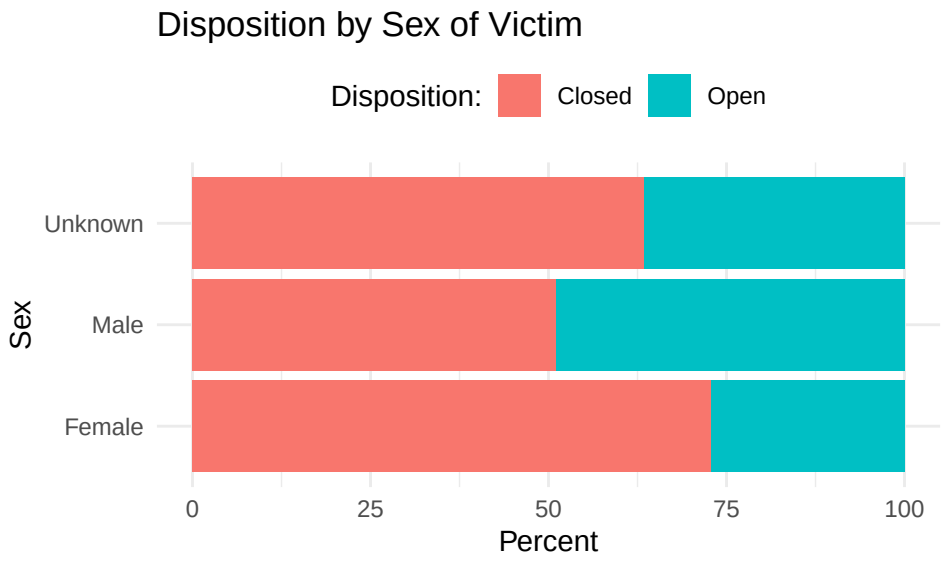



```
crime1 %>% count(sex, disposition.binary) %>% group_by(sex) %>% mutate(Proportion=n/sum(n))
```

```
## # A tibble: 6 x 4
## # Groups:   sex [3]
```

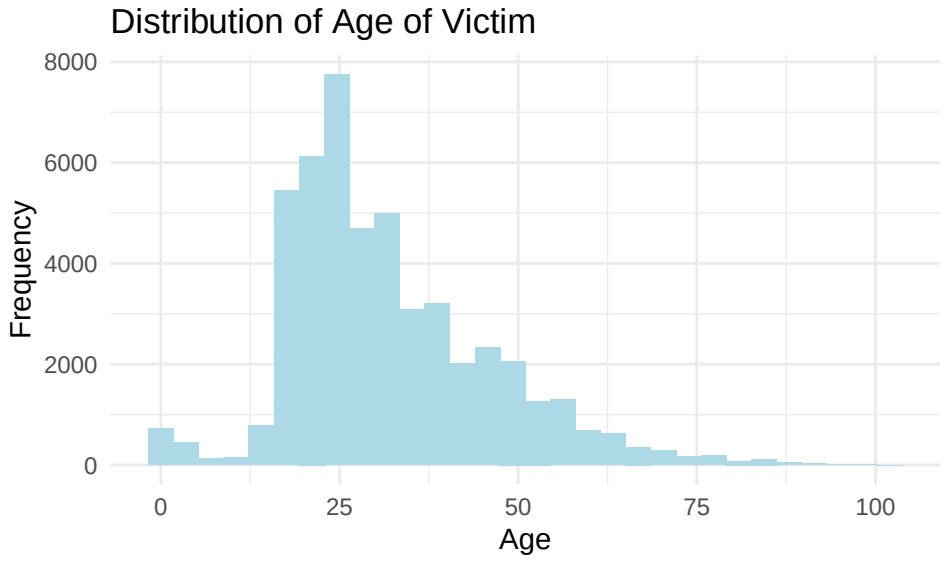
```
##   sex     disposition.binary     n Proportion
##   <fct>   <ord>               <int>      <dbl>
## 1 Female  Open                 1936      0.272
## 2 Female  Closed               5177      0.728
## 3 Male    Open                19782      0.489
## 4 Male    Closed              20663      0.511
## 5 Unknown Open                  593      0.366
## 6 Unknown Closed                1029      0.634
```

```
crime1 %>% count(sex, disposition.binary) %>% group_by(sex) %>% mutate(Proportion=n/sum(n)) %>%
  ggplot(aes(x=sex, y=100*Proportion, fill=disposition.binary)) +
    geom_bar(stat="identity") +
    labs(x="Sex",
         y="Percent",
         title="Disposition by Sex of Victim") +
    theme_minimal() +
  coord_flip() +
  theme(legend.position = "top") +
  scale_fill_discrete(name="Disposition:", limits=c("Closed", "Open"))
```




```
ggplot(data=crime1, aes(x=age)) +
  geom_histogram(fill="lightblue") +
  labs(x="Age",
       y="Frequency",
       title="Distribution of Age of Victim") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



b. Fit a logistic regression model that includes age and sex (with categories female, male, unknown) of the victim. Explain (i.e., interpret) the coefficients for age and sex, and make a graph of the predicted probabilities, with age on the x-axis.

```
fit <- glm(disposition.binary ~ age + sex, family=binomial(link="logit"), data=crime1)
## Note You have to be extra careful what R considers a "success" when you didn't code it as numeric 0 
summary(fit)
```

```
##
## Call:
## glm(formula = disposition.binary ~ age + sex, family = binomial(link = "logit"),
##     data = crime1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7420  -1.1894   0.7967   1.1578   1.2194
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.8310219  0.0343099  24.221  < 2e-16 ***
## age          0.0045229  0.0006471   6.990 2.76e-12 ***
## sexMale     -0.9293664  0.0284851 -32.626  < 2e-16 ***
## sexUnknown  -0.4240415  0.0580738  -7.302 2.84e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 67755  on 49179  degrees of freedom
## Residual deviance: 66458  on 49176  degrees of freedom
## AIC: 66466
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coefficients(fit))
```

```
## (Intercept)         age     sexMale  sexUnknown
```

```
##    2.2956634    1.0045331    0.3948038    0.6543967
```

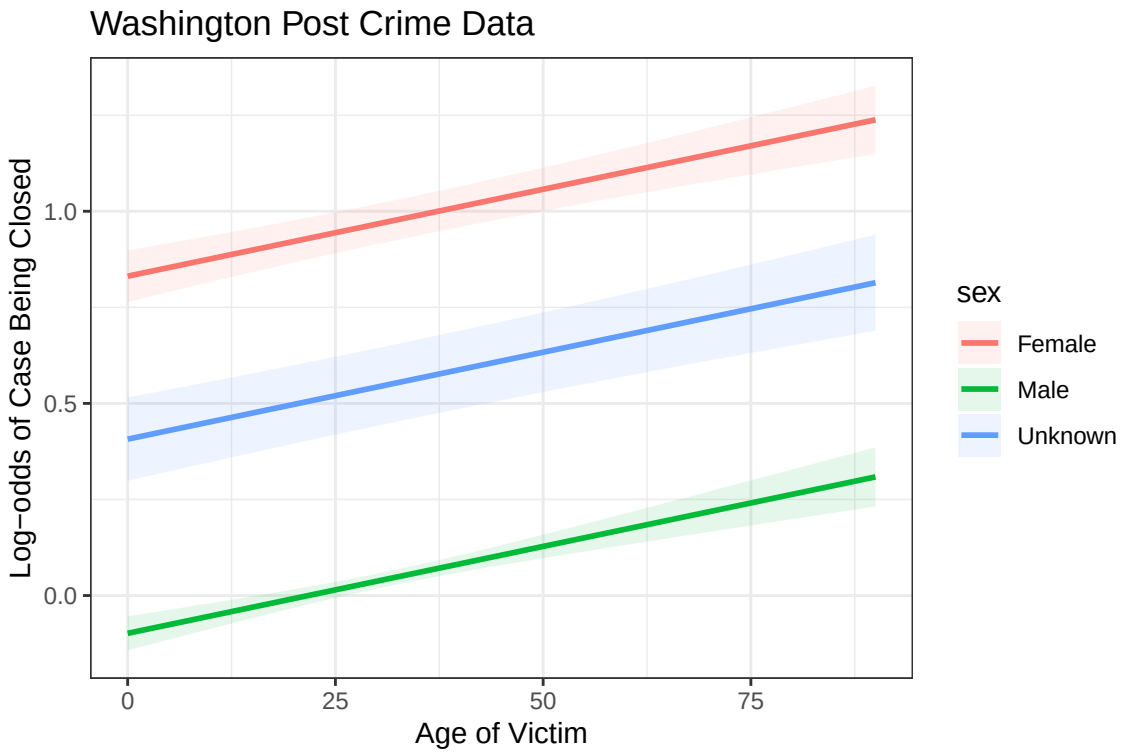
```
exp(10*0.0045229)
```

```
## [1] 1.046267
```

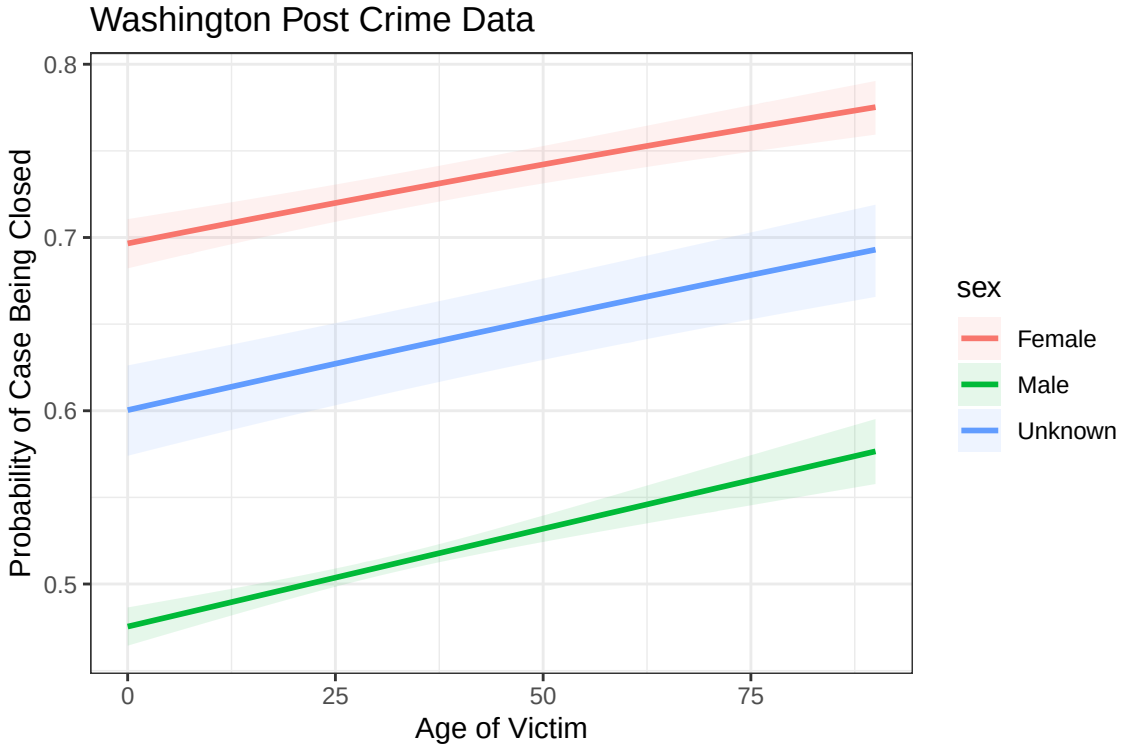
The estimated odds of a case being closed (as opposed to open) increases by 4.6% for every 10 years that a victim is older, controlling for the sex of the victim.

For male victims, the estimated odds of a case being closed are about 60% lower compared to female victims. For victims of unknown sex, the estimated odds of a case being closed are about 35% lower compared to female victims.

```
mygrid <- expand.grid(age = seq(0,90), sex=levels(crime1$sex)) #important to name variables age and sex

predictions <- predict(fit, newdata=mygrid, type="link", se.fit=TRUE)

plotdata <- tibble(
  age = mygrid$age,
  sex = mygrid$sex,
  log.odds = predictions$fit,
  log.odds.LB = log.odds - 1.96*predictions$se.fit,
  log.odds.UB = log.odds + 1.96*predictions$se.fit,
  prob = exp(log.odds)/(1+exp(log.odds)),
  prob.LB = exp(log.odds.LB)/(1+exp(log.odds.LB)),
  prob.UB = exp(log.odds.UB)/(1+exp(log.odds.UB))
)

ggplot(data=plotdata, aes(x=age, y=log.odds, color=sex)) +
  geom_line(size=1) +
  geom_ribbon(aes(ymin=log.odds.LB, ymax=log.odds.UB, fill=sex), color=NA, alpha=0.1) +
  theme_bw() +
  labs(title="Washington Post Crime Data", x="Age of Victim", y="Log-odds of Case Being Closed")
```




```
ggplot(data=plotdata, aes(x=age, y=prob, color=sex)) +
  geom_line(size=1) +
  geom_ribbon(aes(ymin=prob.LB, ymax=prob.UB, fill=sex), color=NA, alpha=0.1) +
  theme_bw() +
  labs(title="Washington Post Crime Data", x="Age of Victim", y="Probability of Case Being Closed")
```



c. Report on the result of a statistical test that tests whether sex is needed in the model.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
Anova(fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: disposition.binary
##     LR Chisq Df Pr(>Chisq)
## age    49.04  1  2.502e-12 ***
## sex  1213.98  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test for sex (likelihood ratio test statistic: 1213, df=2) has an extremely small P-value, indicating that sex is needed in the model.

d. Fit the logistic regression model that includes the interaction between age and sex. Report on the result of a statistical test that tests whether this interaction is needed.

```
fit1 <- glm(disposition.binary ~ age + sex + age*sex, family=binomial(link="logit"), data=crime1)
summary(fit1)
```

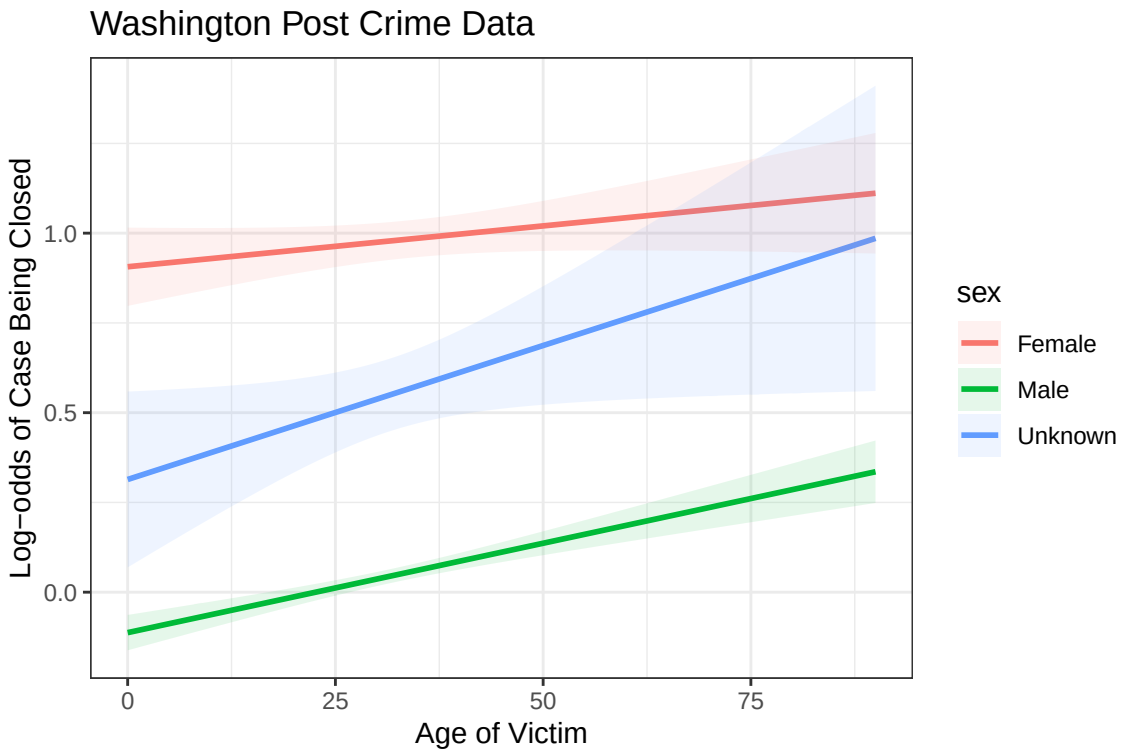
```
##
## Call:
## glm(formula = disposition.binary ~ age + sex + age * sex, family = binomial(link = "logit"),
##     data = crime1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6780  -1.1888   0.7971   1.1577   1.2256
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.906439   0.055612  16.299  < 2e-16 ***
## age             0.002275   0.001447   1.572    0.116
## sexMale        -1.019083   0.061048 -16.693  < 2e-16 ***
## sexUnknown     -0.592511   0.136786  -4.332 1.48e-05 ***
## age:sexMale     0.002704   0.001624   1.664    0.096 .
## age:sexUnknown  0.005188   0.003886   1.335    0.182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 67755  on 49179  degrees of freedom
## Residual deviance: 66455  on 49174  degrees of freedom
## AIC: 66467
##
## Number of Fisher Scoring iterations: 4
```

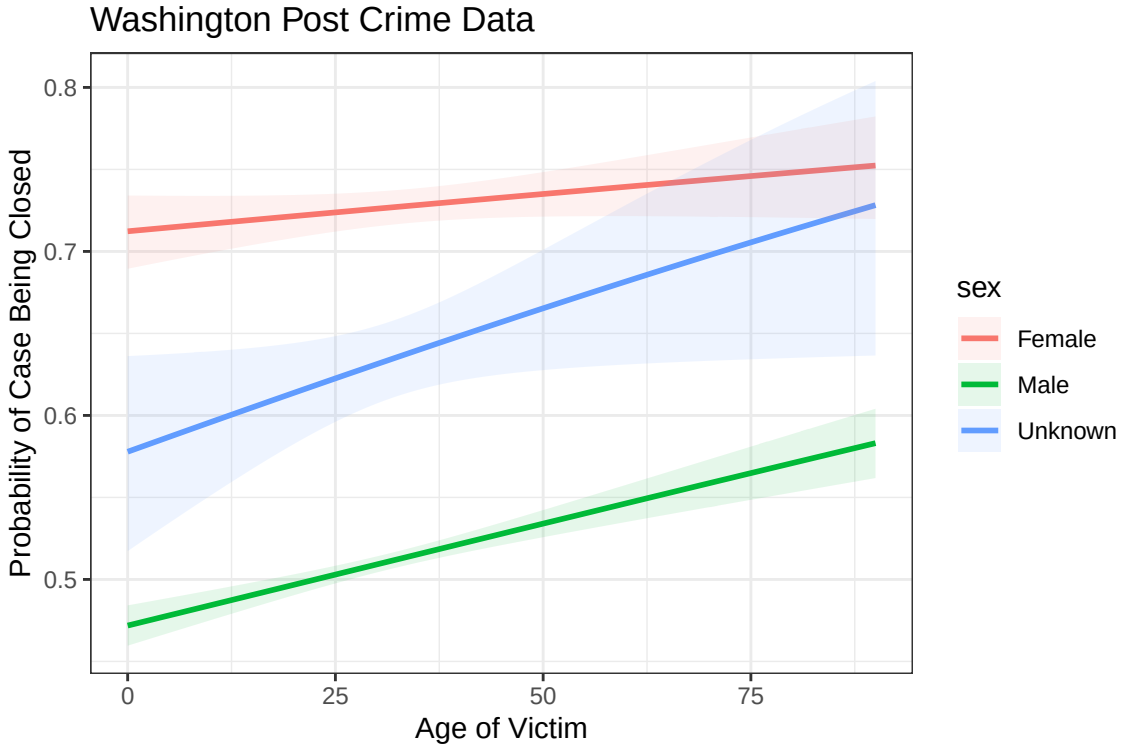
```
Anova(fit1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: disposition.binary
##         LR Chisq Df Pr(>Chisq)
## age        49.04  1  2.502e-12 ***
## sex      1213.98  2  < 2.2e-16 ***
## age:sex     3.45  2     0.1779
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
predictions <- predict(fit1, newdata=mygrid, type="link", se.fit=TRUE)

plotdata <- tibble(
  age = mygrid$age,
  sex = mygrid$sex,
  log.odds = predictions$fit,
  log.odds.LB = log.odds - 1.96*predictions$se.fit,
  log.odds.UB = log.odds + 1.96*predictions$se.fit,
  prob = exp(log.odds)/(1+exp(log.odds)),
  prob.LB = exp(log.odds.LB)/(1+exp(log.odds.LB)),
  prob.UB = exp(log.odds.UB)/(1+exp(log.odds.UB))
)

ggplot(data=plotdata, aes(x=age, y=log.odds, color=sex)) +
  geom_line(size=1) +
  geom_ribbon(aes(ymin=log.odds.LB, ymax=log.odds.UB, fill=sex), color=NA, alpha=0.1) +
  theme_bw() +
  labs(title="Washington Post Crime Data", x="Age of Victim", y="Log-odds of Case Being Closed")
```




```
ggplot(data=plotdata, aes(x=age, y=prob, color=sex)) +
  geom_line(size=1) +
  geom_ribbon(aes(ymin=prob.LB, ymax=prob.UB, fill=sex), color=NA, alpha=0.1) +
  theme_bw() +
  labs(title="Washington Post Crime Data", x="Age of Victim", y="Probability of Case Being Closed")
```



The likelihood ratio test for the age by sex interaction (likelihood ratio test statistic: 3.45, df=2) has P-value = 0.1779. There is no evidence of an age by sex interaction.

e. Plot the ROC curve and find the area under the curve for the model in part b.

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
cutoff <- 0.50
truth <- crime1$disposition.binary
prediction <- factor(fitted(fit) > cutoff, labels = c("Open", "Closed"))
summary(prediction)
```

```
##   Open Closed
##   9580  39600
```

```
confus <- confusionMatrix(data = prediction, reference = truth, positive = "Closed")
confus
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  Open Closed
##     Open    4521   5059
##     Closed 17790  21810
##
##                Accuracy : 0.5354
##                  95% CI : (0.531, 0.5398)
##     No Information Rate : 0.5463
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0151
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8117
##             Specificity : 0.2026
##          Pos Pred Value : 0.5508
##          Neg Pred Value : 0.4719
##              Prevalence : 0.5463
##          Detection Rate : 0.4435
##    Detection Prevalence : 0.8052
##       Balanced Accuracy : 0.5072
##
##        'Positive' Class : Closed
##
```

```
addmargins(confus$table)
```

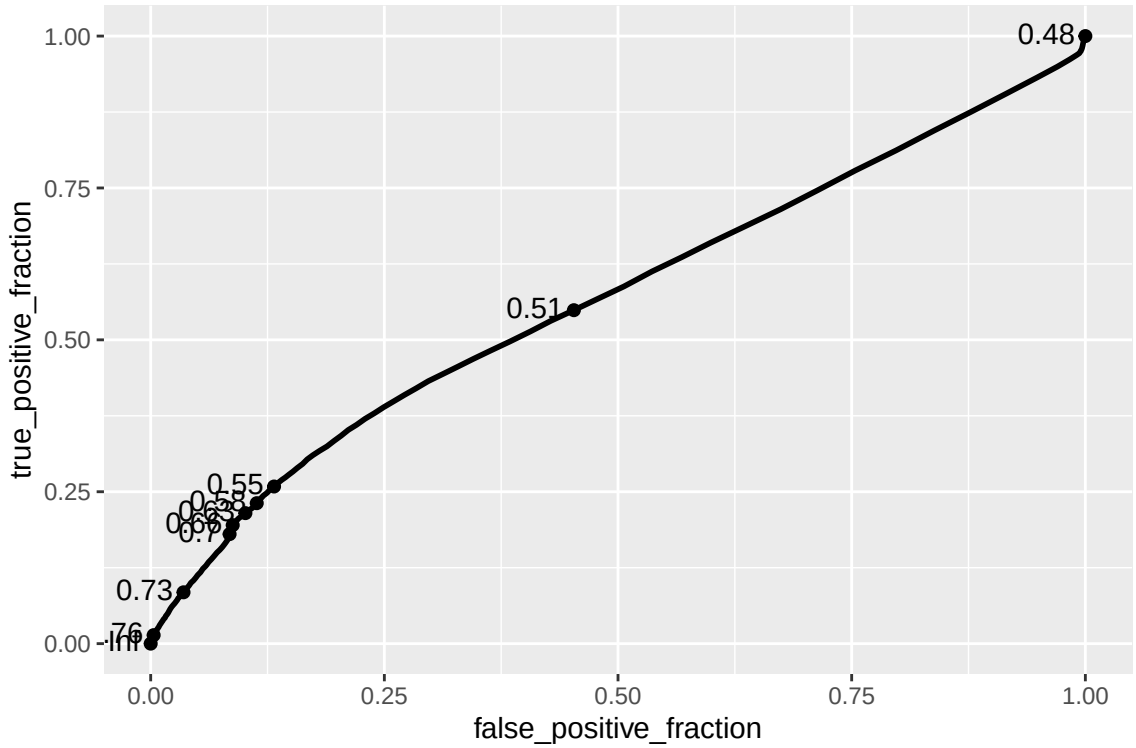
```
##           Reference
## Prediction  Open Closed   Sum
```

```
##      Open    4521   5059  9580
##      Closed 17790  21810 39600
##      Sum    22311  26869 49180
```

```
library(plotROC)
df <- data.frame(response = as.numeric(truth=="Closed"), fit = fitted(fit))
head(df)
```

```
##   response       fit
## 1        1 0.5632693
## 2        1 0.4946364
## 3        1 0.7107163
## 4        1 0.5115949
## 5        1 0.7607272
## 6        0 0.7760172
```

```
ROC.fit2 <- ggplot(df, aes(d = response, m = fit)) +
  geom_roc(labelround = 2)
ROC.fit2
```




```
calc_auc(ROC.fit2)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5681057
```