Logistic Regression

Dav King

3/31/22

Introduction

Multiple regression allows us to relate a numerical response variable to one or more numerical or categorical predictors.

We can use multiple regression models to understand relationships, assess differences, and make predictions.

But what about a situation where the response of interest is categorical and binary?

- spam or not spam
- malignant or benign tumor
- survived or died
- admitted or denied
- won or lost an election

Today's Data: A Night to Remember

On April 15, 1912 the famous ocean liner *Titanic* sank in the North Atlantic after striking an iceberg on its maiden voyage. The dataset `titanic.csv` contains the survival status and other attributes of individuals on the titanic.

- `survived`: survival status (0 = died, 1 = survived)
- `pclass`: passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
- `name`: name of individual
- `sex`: sex (male or female)
- `age`: age in years
- `fare`: passenger fare in British pounds

We are interested in investigating the variables that contribute to passenger survival. Do women and children really come first?

Data and Packages

Today we're using a variety of packages we've used before.

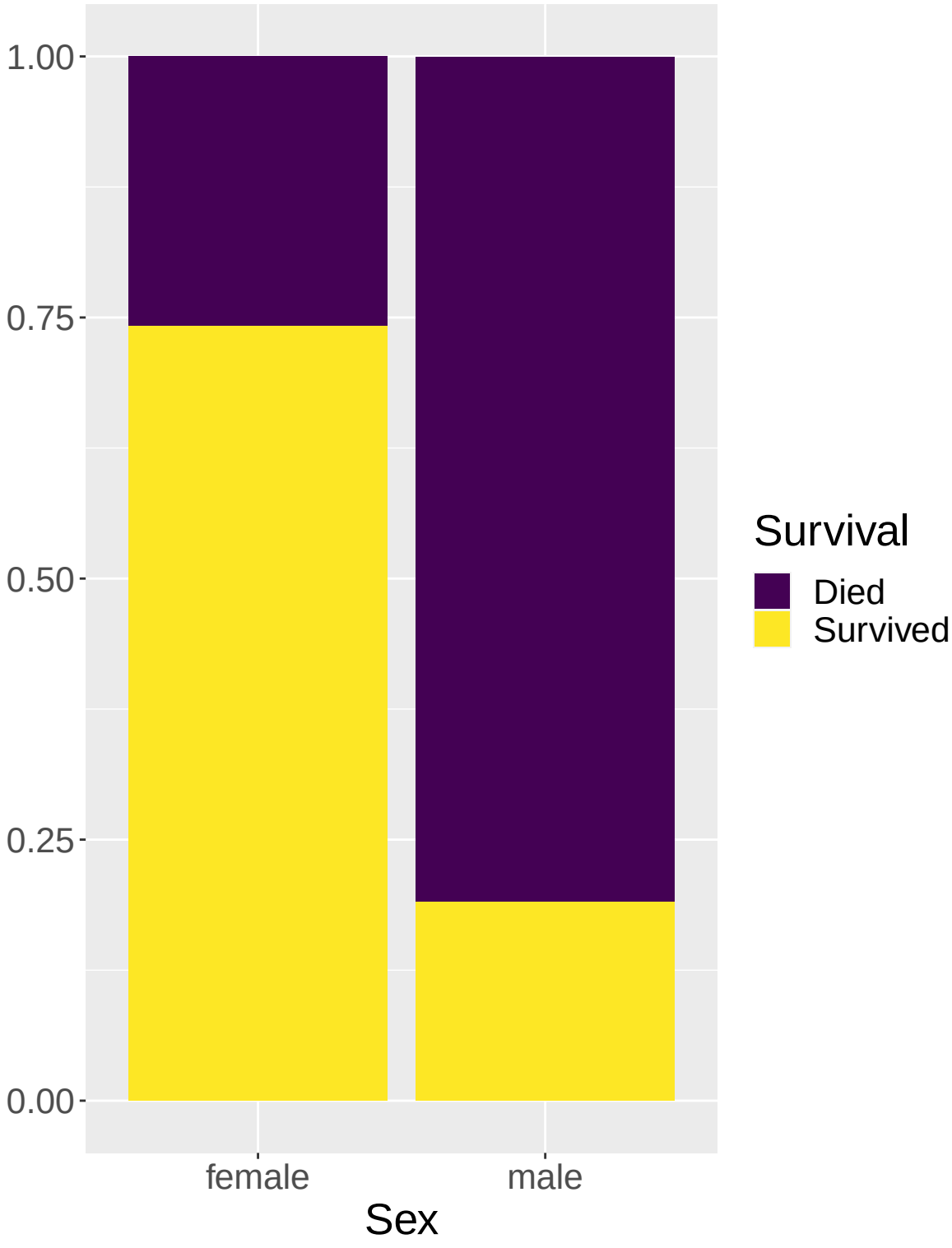
Let's load our data and then look at it.

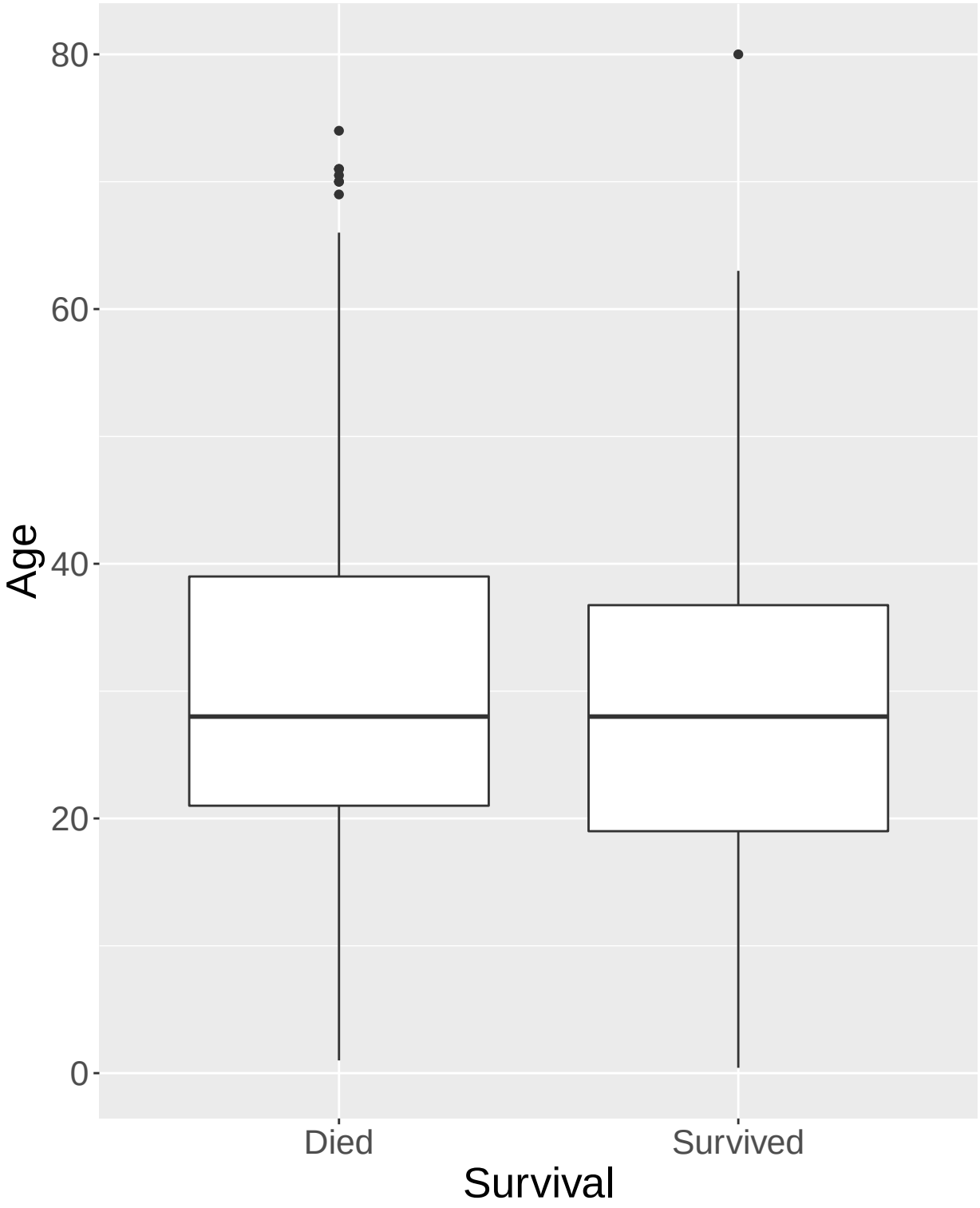
```
titanic <- read_csv("titanic.csv")
```

```
glimpse(titanic)
```

```
## Rows: 887
## Columns: 6
## $ pclass   <dbl> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3, 2~
## $ name     <chr> "Mr. Owen Harris Braund", "Mrs. John Bradley (Florence Briggs~
## $ sex      <chr> "male", "female", "female", "female", "male", "male", "male",~
## $ age      <dbl> 22, 38, 26, 35, 35, 27, 54, 2, 27, 14, 4, 58, 20, 39, 14, 55,~
## $ fare     <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625, 21~
## $ survived <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0~
```

Exploratory Data Analysis





The linear model with multiple predictors

- Population model:

$$y = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \cdots + \beta_k\ x_k + \epsilon$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1\ x_1 + b_2\ x_2 + \cdots + b_k\ x_k$$

Denote by p the probability of death and consider the model below.

$$p = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \cdots + \beta_k\ x_k + \epsilon$$

Can you see any problems with this approach?

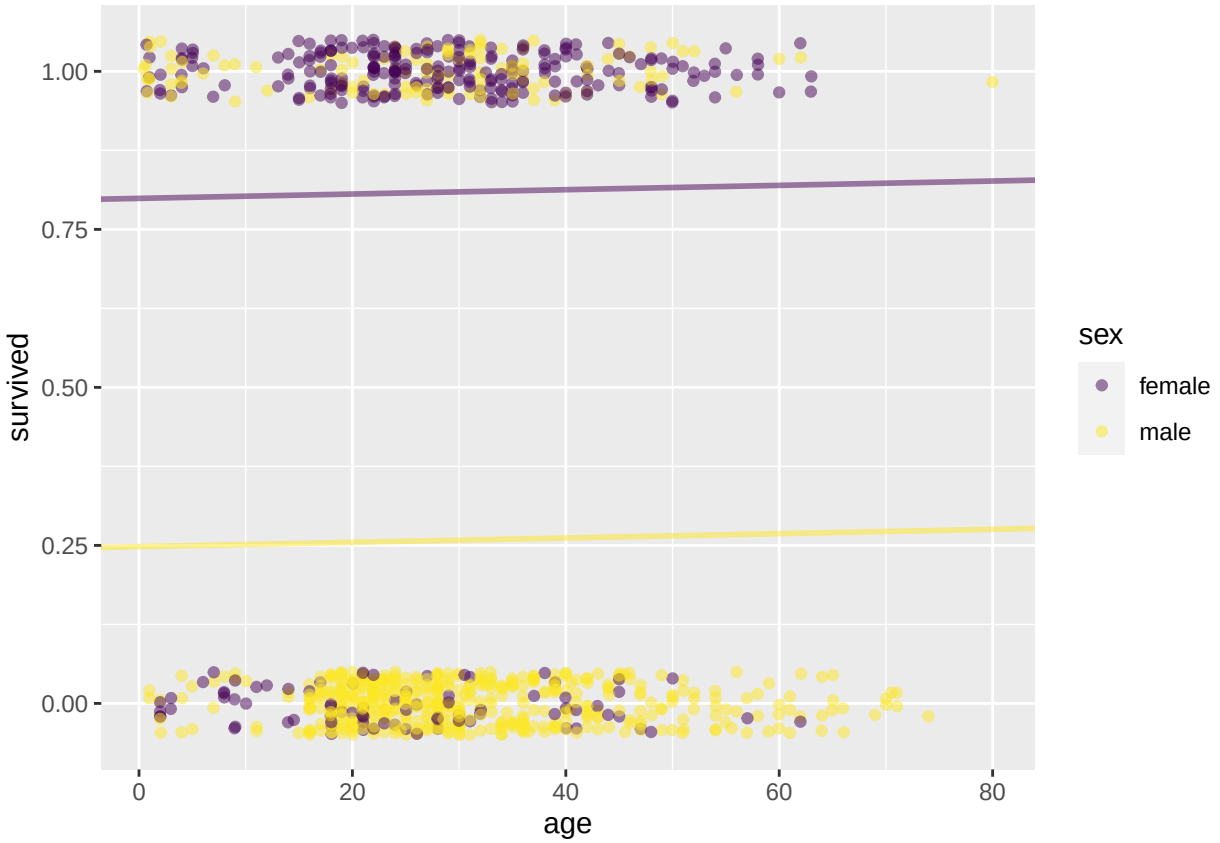
There's no room for nuance - depending on whether something is true or false, its contribution to the prediction is either 0 or 1. While this gives us essentially what a t-test would, it doesn't tell us anything about prediction.

Linear Regression?

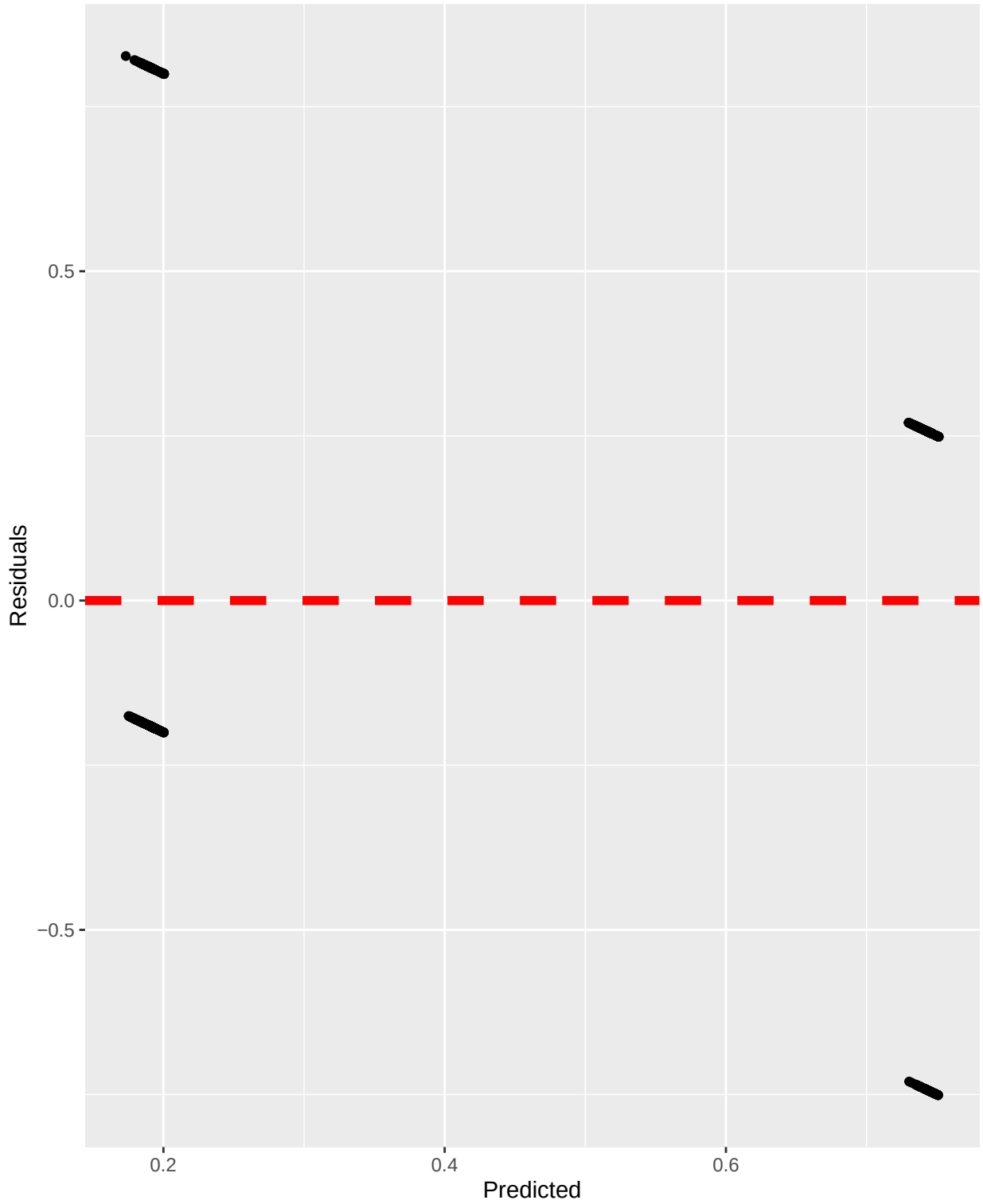
```
lm_survival <- lm(survived ~ age + sex, data = titanic)
tidy(lm_survival)
```

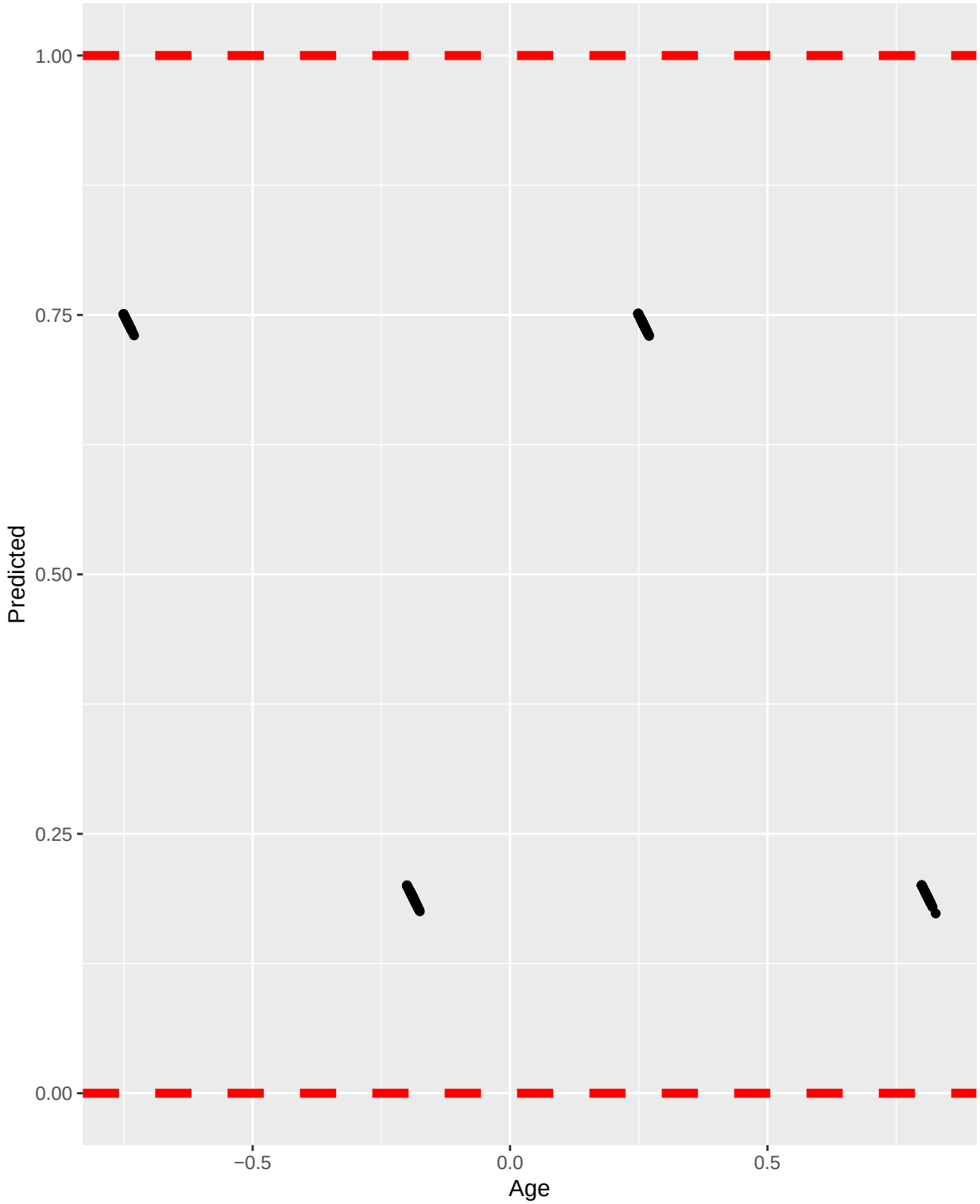
```
## # A tibble: 3 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.752     0.0356      21.1   2.88e-80
## 2 age         -0.000343  0.000979    -0.350 7.26e- 1
## 3 sexmale     -0.551     0.0289     -19.1   3.50e-68
```

Visualizing the Model



Diagnostics





This isn't helpful! We need to develop a new tool.

Preliminaries

- Denote by p the probability of some event
- The **odds** the event occurs is $\frac{p}{1-p}$

Odds are sometimes expressed as X : Y and read X to Y.

It is the ratio of successes to failures, where values larger than 1 favor a success and values smaller than 1 favor a failure.

If $P(A) = 1/2$, what are the odds of A?

1

If $P(B) = 1/3$ what are the odds of B?

0.5

An **odds ratio** is a ratio of odds.

More Preliminaries

- Taking the natural log of the odds yields the **logit** of p

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The logit takes a value of p between 0 and 1 and outputs a value between $-\infty$ and ∞.

The inverse logit (logistic) takes a value between $-\infty$ and ∞ and outputs a value between 0 and 1.

$$\text{inverse logit}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

There is a one-to-one relationship between probabilities and log-odds. If we create a model using the log-odds we can "work backwards" using the logistic function to obtain probabilities between 0 and 1.

Logistic Regression Model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Use the inverse logit to find the expression for p.

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}$$

We can use the logistic regression model to obtain predicted probabilities of success for a binary response variable.

Logistic Regression Model

We can handle fitting the model via computer either in a `tidymodels` framework or by using the `glm` function.

```
fit_1 <- logistic_reg() %>%
  set_engine("glm") %>%
  fit(as.factor(survived) ~ sex + age, data = titanic, family = "binomial")

fit_1 %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  1.11      0.208       5.34  9.05e- 8
## 2 sexmale     -2.50      0.168     -14.9   3.24e-50
## 3 age         -0.00206   0.00586    -0.351 7.25e- 1
```

```
logit_mod <- glm(survived ~ sex + age, data = titanic, family = "binomial")
tidy(logit_mod)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  1.11      0.208       5.34  9.05e- 8
## 2 sexmale     -2.50      0.168     -14.9   3.24e-50
## 3 age         -0.00206   0.00586    -0.351 7.25e- 1
```

And use `augment` to find predicted log-odds.

```
pred_log_odds <- augment(logit_mod)
```

The Estimated Logistic Regression Model

```
tidy(logit_mod)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  1.11      0.208       5.34  9.05e- 8
## 2 sexmale     -2.50      0.168     -14.9   3.24e-50
## 3 age         -0.00206   0.00586    -0.351 7.25e- 1
```

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.11 - 2.50\ sex - 0.00206\ age$$

$$\hat{p} = \frac{e^{1.11-2.50\ sex-0.00206\ age}}{1+e^{1.11-2.50\ sex-0.00206\ age}}$$

Interpreting coefficients

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.11 - 2.50\ sex - 0.00206\ age$$

Holding sex constant, for every additional year of age, we expect the log-odds of survival to decrease by approximately 0.002.

Holding age constant, we expect males to have a log-odds of survival that is 2.50 less than females.

Interpreting coefficients

$$\frac{\hat{p}}{1-\hat{p}} = e^{1.11 - 2.50\ sex - 0.00206\ age}$$

```
tidy(logit_mod) %>%
  mutate(estimate= exp(estimate))
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   3.05     0.208       5.34  9.05e- 8
## 2 sexmale       0.0821   0.168     -14.9   3.24e-50
## 3 age           0.998    0.00586    -0.351 7.25e- 1
```

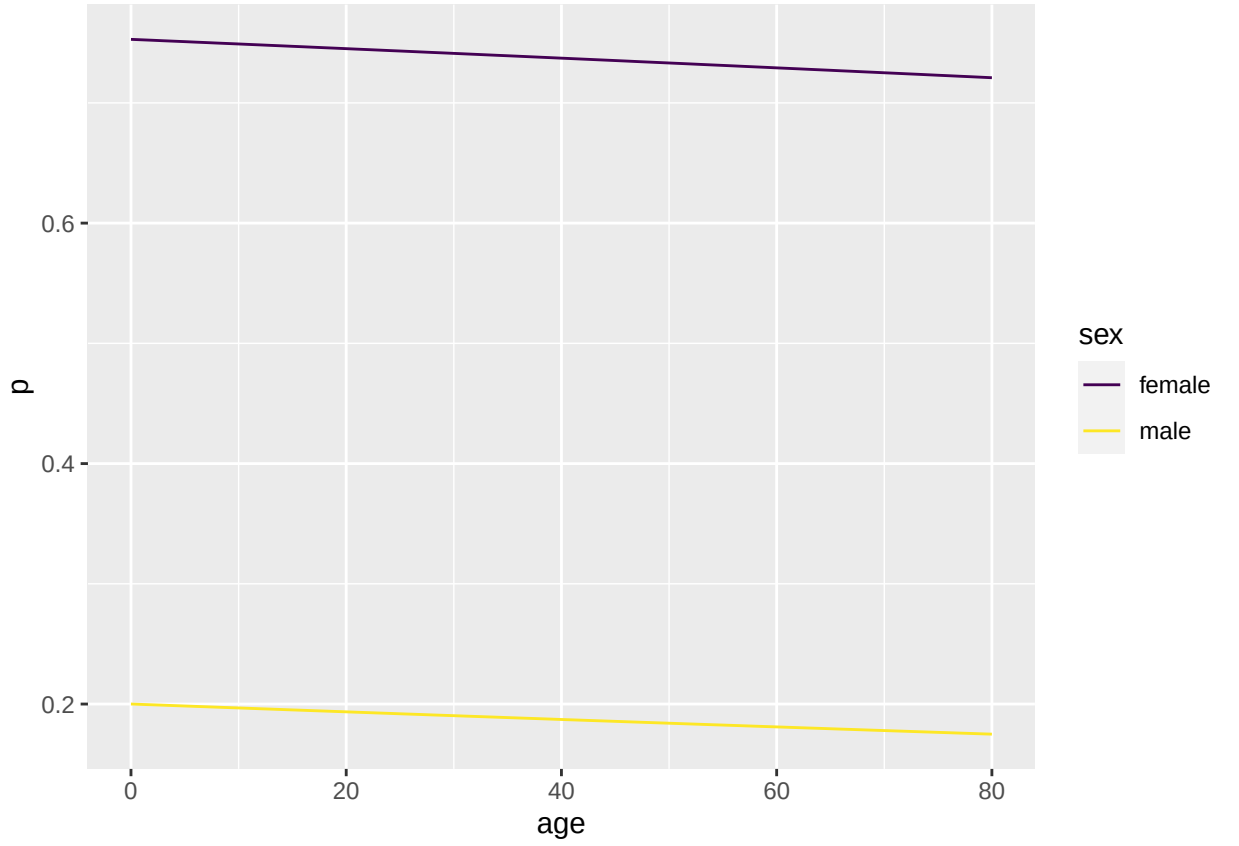
Holding sex constant, for every one year increase in age, the odds of survival is expected to be multiplied by $e^{-0.00206} = 0.998$.

Holding age constant, the odds of survival for males is $e^{-2.50} = 0.082$ times the odds of survival for females.

Predicted Probabilities

- We can also obtain the predicted probability of success at different levels of the explanatory variables and then plot these probabilities.

```
tibble(
  age = rep(0:80, times = 2),
  sex = rep(c("male", "female"), each = 81)
  ) %>%
  augment(logit_mod, newdata = .) %>%
  mutate(p = exp(.fitted) / (1 + exp(.fitted))) %>%
  ggplot(aes(x = age, y = p, color = sex)) +
  geom_line() +
  scale_color_viridis_d()
```



- **Question**: What do you notice about the effect of age and sex here?

Sex predicts a huge discrepancy in probability of survival - dropping it by over 0.7 across the board. Age predicts a slight decrease in survival over time. Interestingly, there is an interaction effect between these - the effect of age on survival probability varies depending on sex.

Specific probabilities

- Let's say you're interested in the probability of survival of someone of a specific age and sex. You can also calculate this to get a specific number rather than including a plot.

```
tibble(
  age = 31,
  sex = "male") %>%
  augment(logit_mod, newdata = .) %>%
  mutate(p = exp(.fitted) / (1 + exp(.fitted))) %>%
  pull("p")
```

```
## [1] 0.1900059
```

Model Fit

```
glance(logit_mod)
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1         1183.     886  -458.  922.  936.     916.         884   887
```

- If you use `glance`, you can see a variety of measures of model fit.
- Two measures you can see are the null deviance and the deviance. The null deviance tells us how well we can predict the response variable with only the intercept, while the deviance tells us about the model fit now that we added two predictors. Notice that the deviance drops from 1183 to 916, with the loss of two degrees of freedom for the two predictors we added to the model.
- Here, lower values are better. There is a reduction in the deviance by 267 with a loss of two degrees of freedom.
- You can also see the AIC (and BIC) here. This is based upon the deviance, but penalizes you for including more explanatory variables, like we saw for adjusted R^2. AIC is useful when comparing different models.

Weaknesses

- Logistic regression has assumptions: independence and linearity in the log-odds (some other methods require fewer assumptions)
- If the predictors are correlated, coefficient estimates may be unreliable

Strengths

- Can transform to odds ratios or predicated probabilities for interpretation of coefficients.
- Handles numerical and categorical predictors
- Can quantify uncertainty around a prediction
- Can extend to more than 2 categories (multinomial regression)

Practice Problems

1. Please add fare to the model. Interpret the coefficients for your variables using odds ratios.

```
fare_mod <- glm(survived ~ sex + age + fare, data = titanic)
out <- tidy(fare_mod) %>%
  mutate(estimate = exp(estimate))
out
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    2.01   0.0361       19.4  5.08e-70
## 2 sexmale        0.596  0.0289      -17.9  1.64e-61
## 3 age            0.999  0.000969     -1.13 2.59e- 1
## 4 fare           1.00   0.000278      5.89 5.50e- 9
```

Holding age and fare constant, the odds of survival of males is 0.596 times the odds of survival for females.

Holding sex and fare constant, for every one year increase in age, we expect the odds of survival to be multiplied by 0.999.

Holding sex and age constant, for each additional pound paid in fare we expect the odds of survival to be multiplied by 1.002.

2. What is the predicted probability of survival for a 40 year old man who paid 100 pounds? What if it went up to 500 pounds?

```
tibble(age = 40, sex = "male", fare = 100) %>%
  augment(fare_mod, newdata = .) %>%
  mutate(p = exp(.fitted)/(1 + exp(.fitted))) %>%
  pull("p")
```

```
## [1] 0.5748482
```

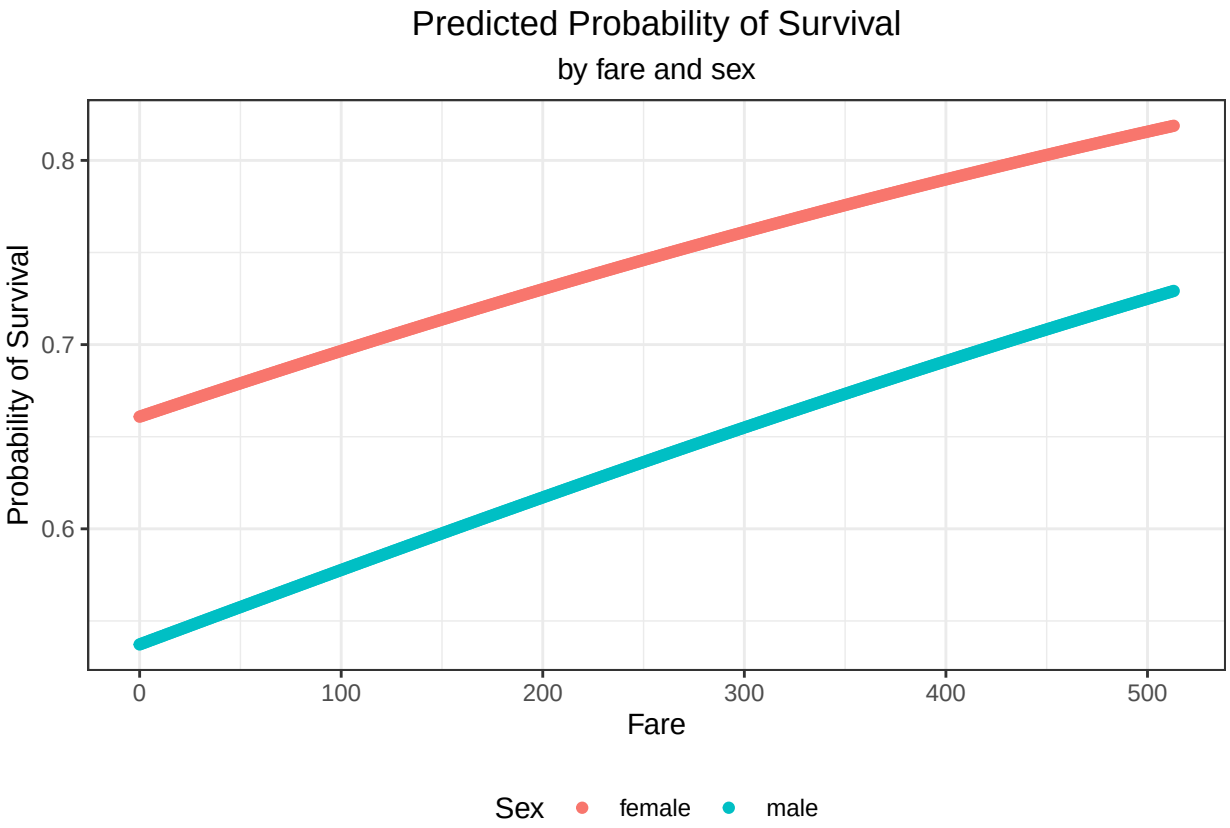
```
tibble(age = 40, sex = "male", fare = 500) %>%
  augment(fare_mod, newdata = .) %>%
  mutate(p = exp(.fitted)/(1 + exp(.fitted))) %>%
  pull("p")
```

```
## [1] 0.7225925
```

The predicted probability of a 40 year old man who paid 100 pounds is 0.575, and if he paid 500 pounds that goes up to 0.723.

3. Set age as being equal to its mean value. Then, create a predicted probability plot showing the effect of fare price for men and women. Describe what you see.

```
tibble(age = mean(titanic$age),
       fare = rep(0:513, times = 2),
       sex = rep(c("male", "female"), each = 514)) %>%
  augment(fare_mod, newdata = .) %>%
  mutate(p = exp(.fitted)/(1 + exp(.fitted))) %>%
  ggplot(aes(x = fare, y = p, color = sex)) +
  geom_point() +
  theme_bw() +
  labs(title = "Predicted Probability of Survival",
       subtitle = "by fare and sex", x = "Fare",
       y = "Probability of Survival", color = "Sex") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom")
```



Though there are clear differences in the probability of survival between the sexes (with women substantially more likely to have survived), the more money that was spent on fare the more likely members of both sexes were to survive.

Sources

- Computing for the Social Sciences. "Logistic Regression." https://cfss.uchicago.edu/notes/logistic-regression/
- Lillis, David. Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output