Logistic regression

Computational Mathematics and Statistics Camp
University of Chicago
September 2018

1. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. *Temp* gives the temperature in Fahrenheit, *Damaged* represents the number of damaged O-rings, and *Undamaged* represents the number of O-rings that were not damaged.

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

   a. Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

   More damaged O-rings are observed for lower temperatures, with fewer damaged O-rings seen for higher temperatures. The lowest-temperature shuttle mission had five damaged O-rings, much more than any other shuttle launch, so this observation will be especially influential on the results of an analysis.

   b. Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below. Describe the key components of this summary table in words.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

   We are generally most interested interested in the coefficients of variables, so in this case, the *Temperature* row. The coefficient of this term is negative, indicating that increasing temperatures are associated with a lower probability of O-ring damage. This coefficient was statistically significant with a p-value near 0 ($H_0 : \beta_{Temp.} = 0$, $H_A : \beta_{Temp.} \neq = 0$), indicating that the data provide strong evidence that the coefficient is less than 0.

   c. Write out the logistic model using the point estimates of the model parameters.

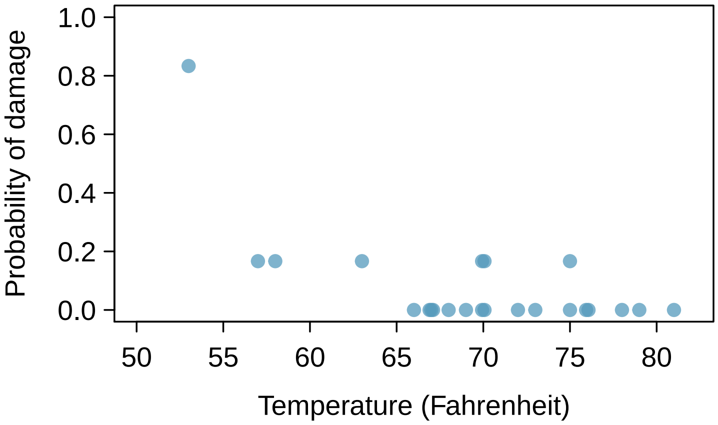
   $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times \text{Temperature}$, where $\hat{p}$ represents the probability of damage to an O-ring.

   d. Based on the model, do you think concerns regarding O-rings are justified? Explain.

   Yes. While the data are observational, the relationship between temperature and damage to

O-rings is very strong. Since lives are at stake, such a strong association indicates something important is going on that must be carefully investigated. (Ultimately, O-rings were cited as the cause for the shuttle disaster, though this investigation required much more investigation than what was completed in this exercise to come to a causal conclusion.)

2. Continuing with the investigation above:



a. The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

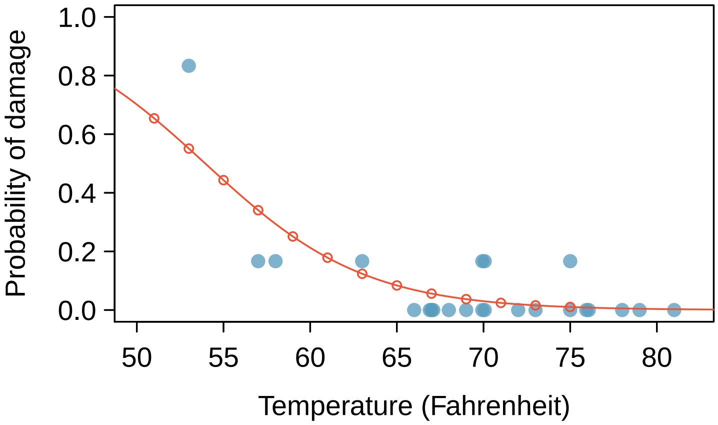
where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged. Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$$\hat{p}_{57} = 0.341 \quad \hat{p}_{59} = 0.251 \quad \hat{p}_{61} = 0.179 \quad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \quad \hat{p}_{67} = 0.056 \quad \hat{p}_{69} = 0.037 \quad \hat{p}_{71} = 0.024$$

The calculation of the probabilities is shown below:

$$\log\left(\frac{\hat{p}_{51}}{1-\hat{p}_{51}}\right) = 11.6630 - 0.2162 \times 51 \to \hat{p}_{51} = \frac{e^{11.6630 = -.2162 \times 51}}{1 + e^{11.6630 = -.2162 \times 51}} = 0.654$$
$$\log\left(\frac{\hat{p}_{53}}{1-\hat{p}_{53}}\right) = 11.6630 - 0.2162 \times 53 \to \hat{p}_{53} = \frac{e^{11.6630 = -.2162 \times 53}}{1 + e^{11.6630 = -.2162 \times 53}} = 0.551$$
$$\log\left(\frac{\hat{p}_{55}}{1-\hat{p}_{55}}\right) = 11.6630 - 0.2162 \times 55 \to \hat{p}_{55} = \frac{e^{11.6630 = -.2162 \times 55}}{1 + e^{11.6630 = -.2162 \times 55}} = 0.443$$

b. Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.



c. Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

For logistic regression to be appropriate in this case, each O-ring must be independent of the others. We must assume independence is true (or very nearly so) if we are to believe the model. For example, if the O-ring manufacturing process has changed dramatically over the time of the different shuttle launches, then we should be skeptical of independence and therefore also of the model.

3. An important question in American politics is why do some people participate in the political process, while others do not? Participation has a direct impact on outcomes – if you fail to participate in politics, the government and political officials are less likely to respond to your concerns. Typical explanations focus on a resource model of participation – individuals with greater resources, such as time, money, and civic skills, are more likely to participate in politics. One area of importance is understanding voter turnout, or why people participate in elections. Using the resource model of participation as a guide, we can develop several expectations. First, women, who more frequently are the primary caregiver for children and earn a lower income, are less likely to participate in elections than men. Second, older Americans, who typically have more time and higher incomes available to participate in politics, should be more likely to participate in elections than younger Americans. Finally, individuals with more years of education, who are generally more interested in politics and understand the value and benefits of participating in politics, are more likely to participate in elections than individuals with fewer years of education.

While these explanations have been repeatedly tested by political scientists, an emerging theory assesses an individual's mental health and its effect on political participation.[1] Depression increases individuals' feelings of hopelessness and political efficacy, so depressed individuals will have less desire to participate in politics. More importantly to our resource model of participation, individuals with depression suffer physical ailments such as a lack of energy, headaches, and muscle soreness which drain an individual's energy and requires time and money to receive treatment. For these reasons, we should expect that individuals with depression are less likely to participate in election than those without symptoms of depression.

The 1998 General Social Survey included several questions about the respondent's mental health, including:

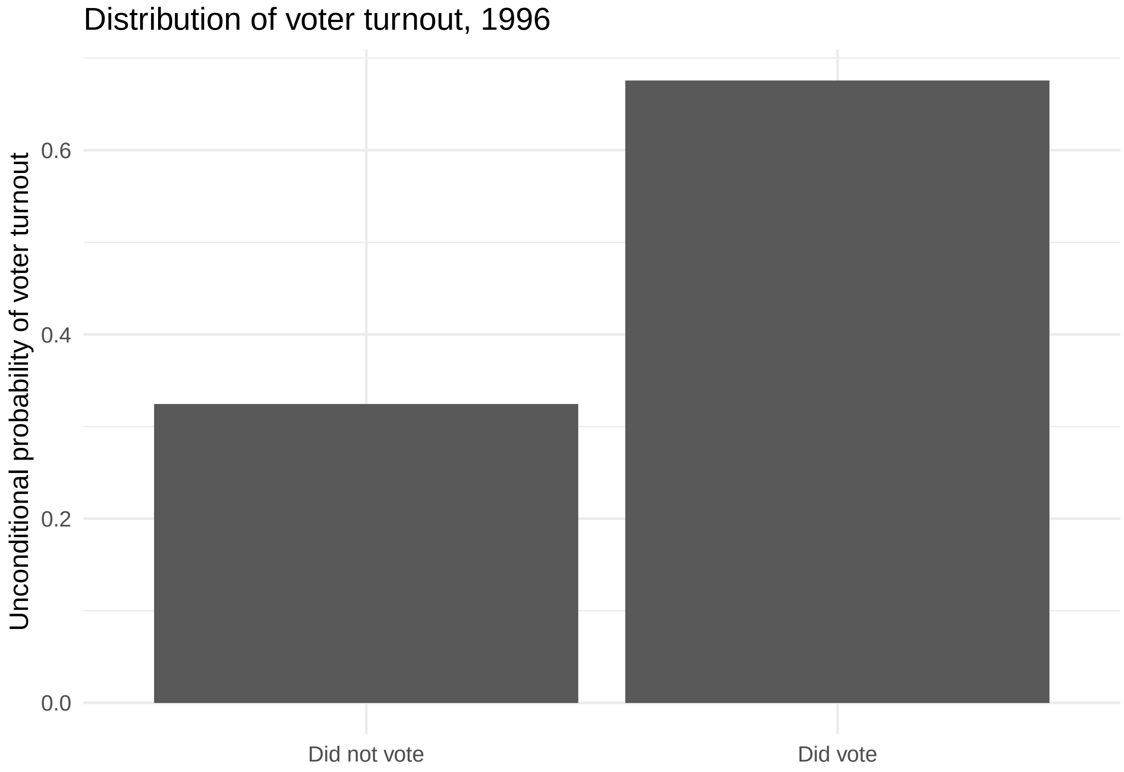
- `vote96` - 1 if the respondent voted in the 1996 presidential election, 0 otherwise
- `mhealth_sum` - index variable which assesses the respondent's mental health, ranging from 0 (an individual with no depressed mood) to 9 (an individual with the most severe depressed mood)[2]

[1] Ojeda, C. (2015). Depression and political participation. *Social Science Quarterly*, 96(5), 1226-1243.

[2] The variable is an index which combines responses to four different questions: "In the past 30 days, how often did you feel: 1) so sad nothing could cheer you up, 2) hopeless, 3) that everything was an effort, and 4) worthless?" Valid responses are none of the time, a little of the time, some of the time, most of the time, and all of the time.

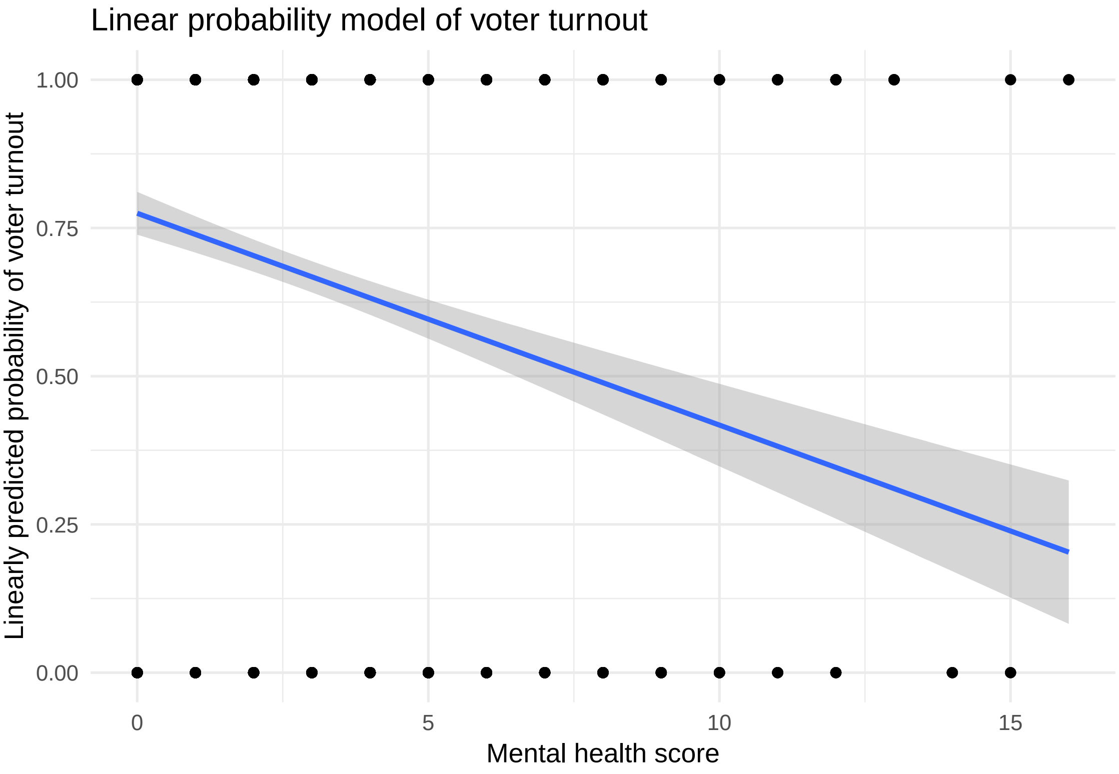
- `black` - 1 if the respondent is black, 0 otherwise
- `female` - 1 if the respondent is female, 0 if male

a. Below is a histogram of voter turnout. What is the approximate unconditional probability of a given individual turning out to vote?



The unconditional probability of voter turnout in 1996 is 0.676.

b. Below is an OLS model of the relationship between mental health score and probability of voter turnout. What information does this tell us? What is problematic about this linear smoothing line?



There appears to be a negative and statistically significant relationship between mental health and voter turnout - as mental health score increases (aka as individuals become more depressed), their probability of turning out to vote decreases. However this is not strictly speaking generating predicted probabilities because the smoothing line assumes a continuous outcome variable, which we do not have here. A logistic regression model with the logit transformed linear predictor is more appropriate given our data.

c. Below are the results of a basic logistic regression model of mental health on the probability of voter turnout.

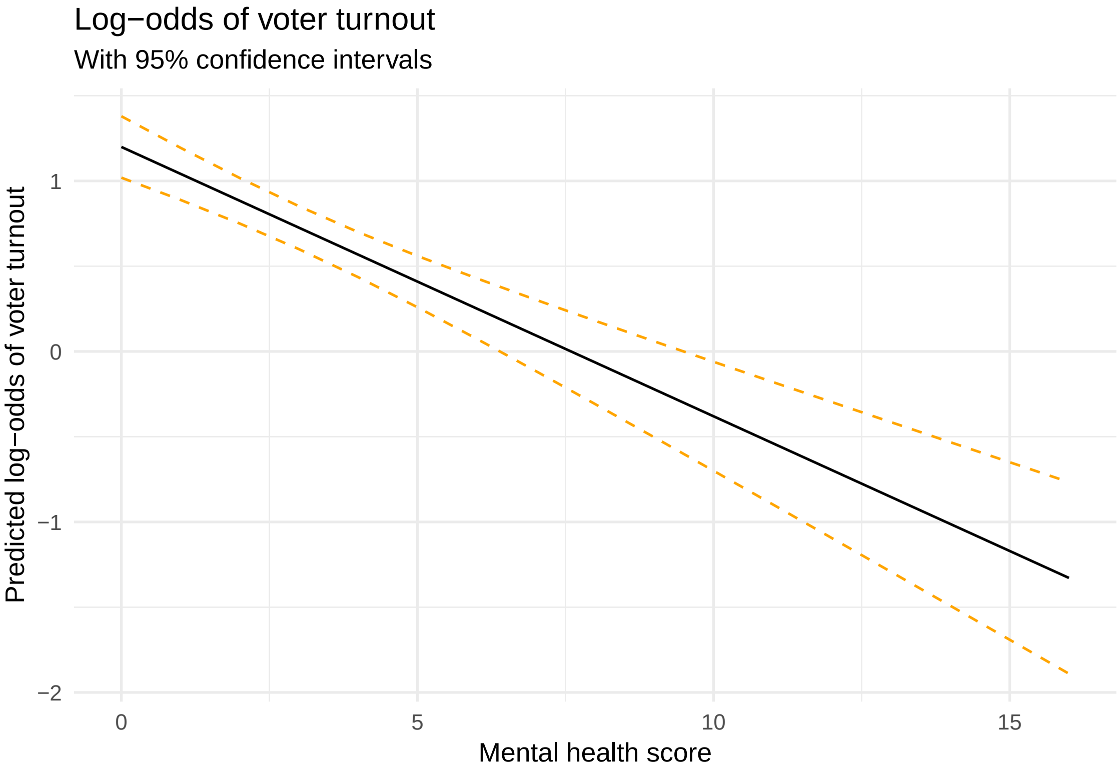
```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum, family = binomial, data = mh)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.710  -1.286   0.726   0.832   1.768
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1994     0.0919   13.05  < 2e-16 ***
## mhealth_sum  -0.1580     0.0216   -7.32  2.4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1468.3  on 1164  degrees of freedom
## Residual deviance: 1411.8  on 1163  degrees of freedom
## AIC: 1416
```

```
##
## Number of Fisher Scoring iterations: 4
```

Is the relationship between mental health and voter turnout statistically and/or substantively significant? Justify your answer.

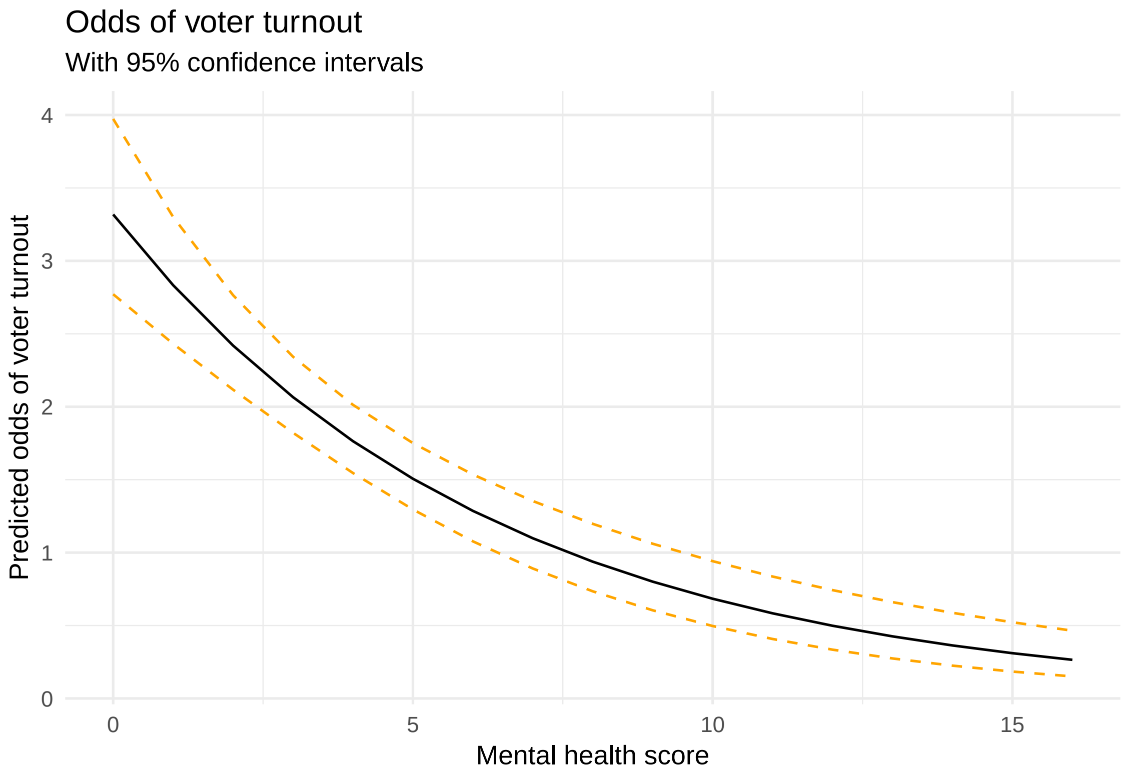
Yes, the relationship appears statistically significant. The p-value is less than .05. The relationship appears negative, so that as depression increases voter turnout also decreases. However the parameter is not easily interpreted in its current log-odds form.

d. Use the previous results table and this graph to interpret the results in terms of log-odds.



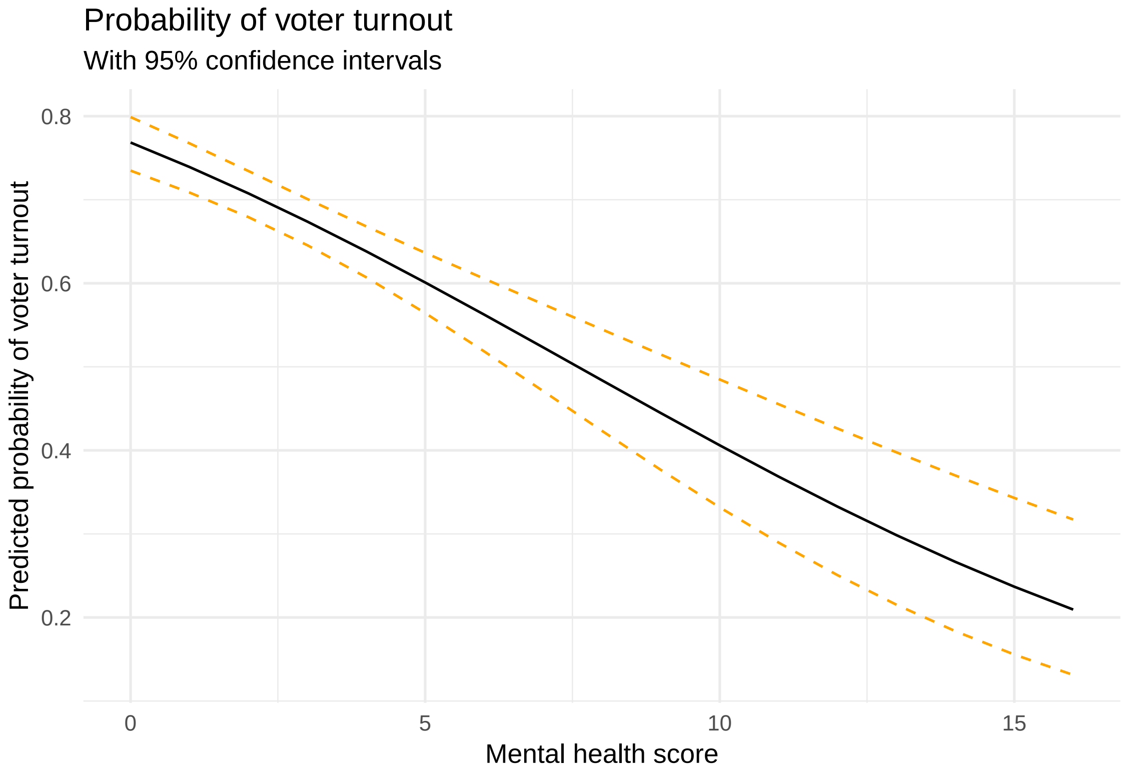
For every one-point increase in depression, the log-odds of voter turnout decrease by 0.158.

e. Use the previous results table and this graph to interpret the results in terms of odds.



For every one-point increase in depression, the predicted odds of voter turnout are 0.854 times the size at the previous level. Aka the odds ratio for depression and voter turnout is 0.854.

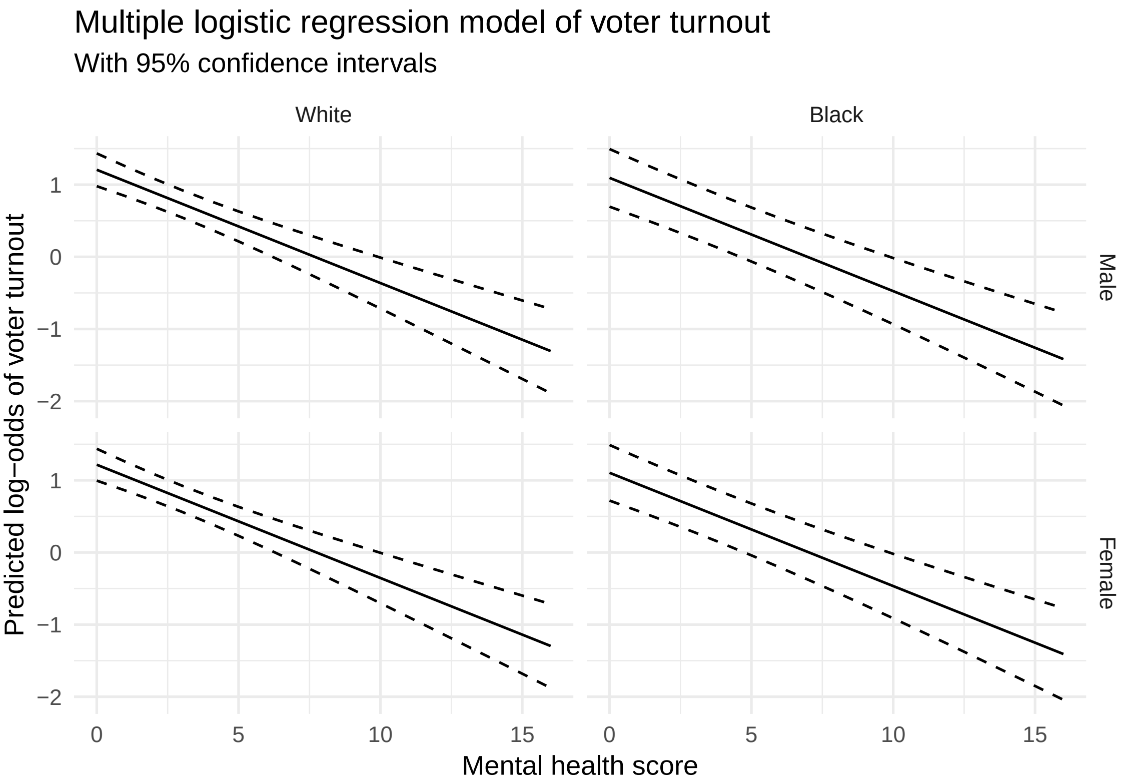
f. Use the previous results table and this graph to interpret the results in terms of probability.

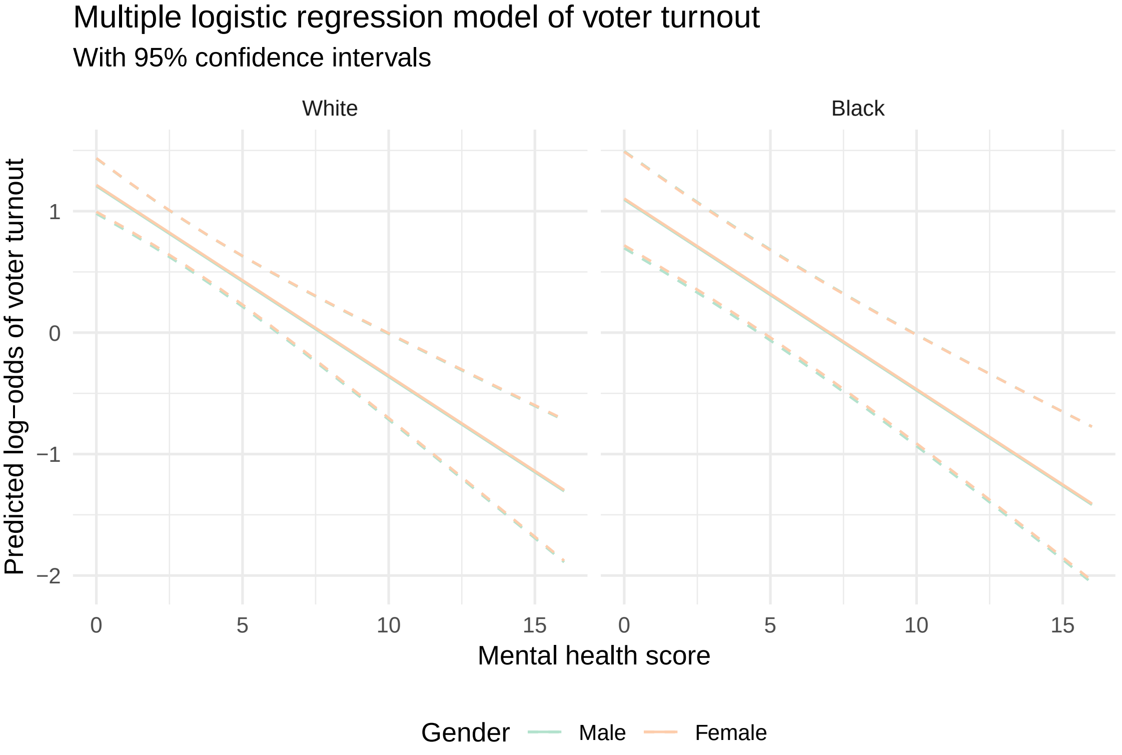


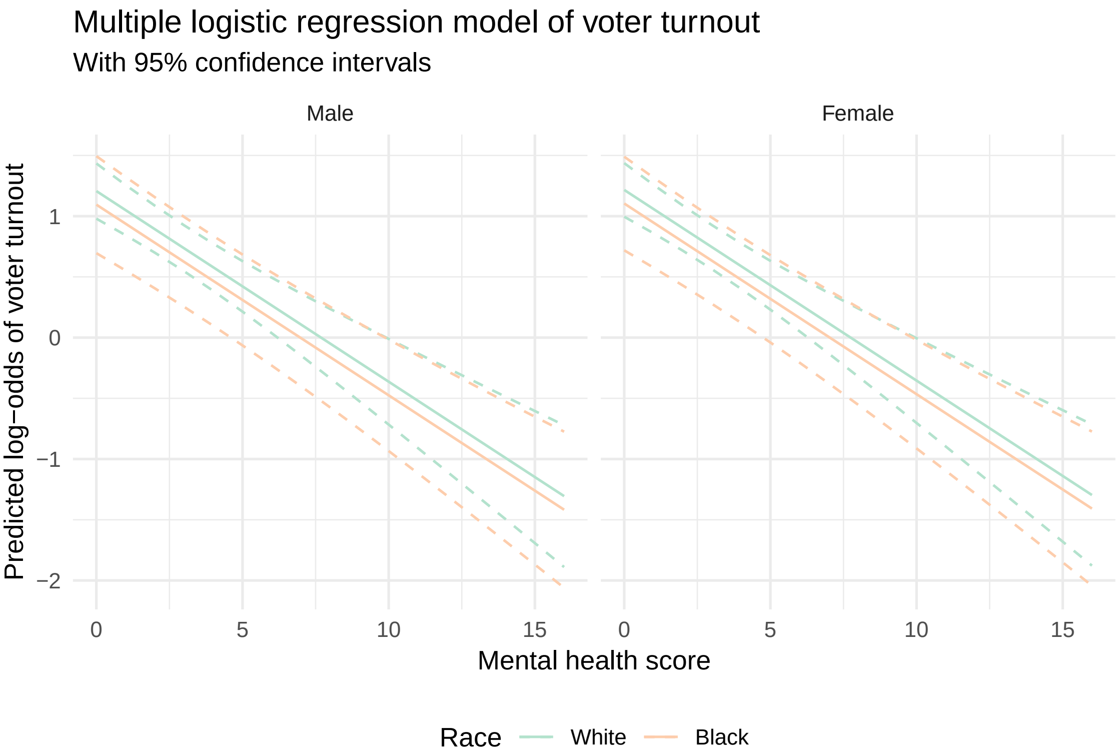
The probability of turning out to vote decreases as one's mental health score increases. This change in probability is larger for those with medium mental health scores (slope is larger) and less so for those with already low or high scores.

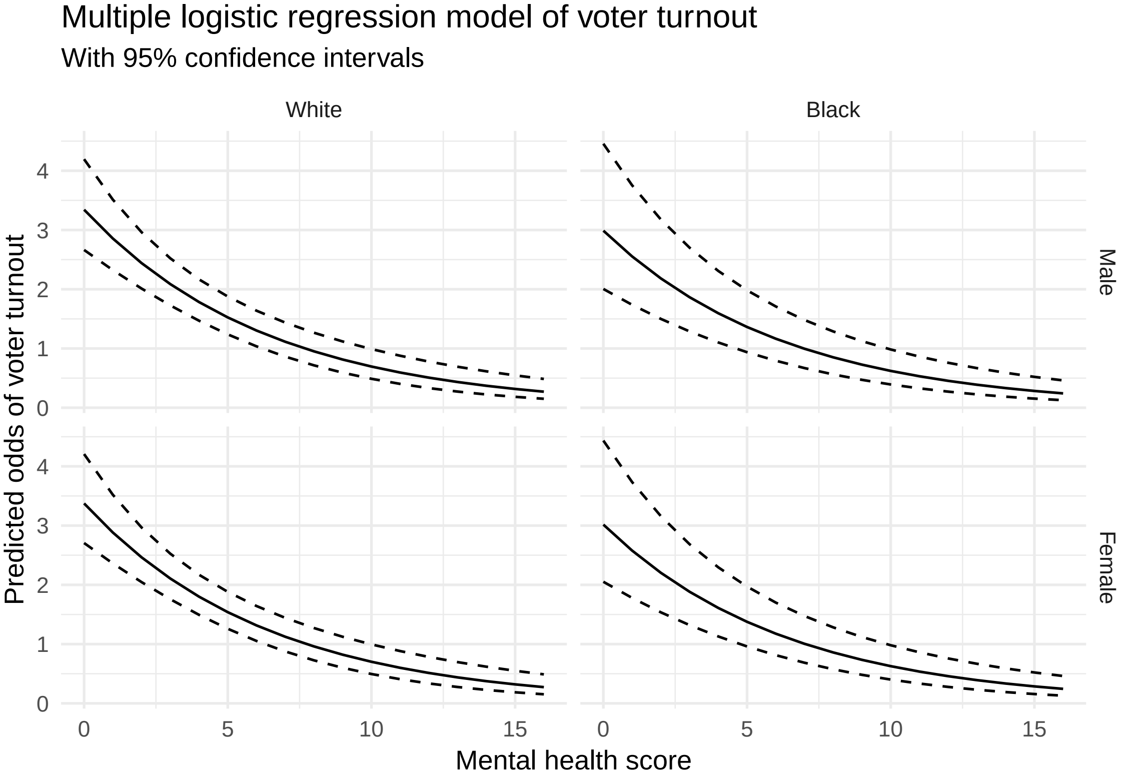
g. Here are the results of a multivariate logistic regression model and some related graphs. Interpret the results in paragraph format. This should include a discussion of your results as if you were reviewing them with fellow computational social scientists. Discuss the results using any or all of log-odds, odds, predicted probabilities, and first differences - choose what makes sense to you and provides the most value to the reader.

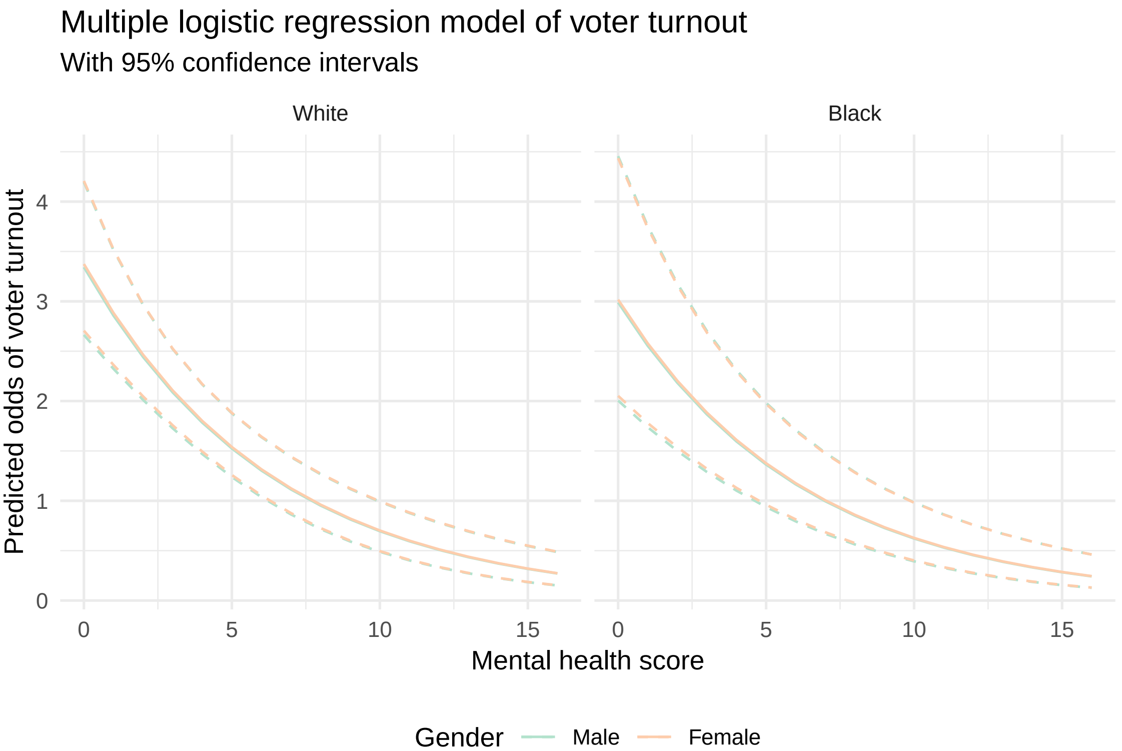
```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum + female + black, family = binomial,
##     data = mh)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.718  -1.292   0.760   0.829   1.808
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.20678    0.11590   10.41  < 2e-16 ***
## mhealth_sum -0.15702    0.02162   -7.26  3.8e-13 ***
## female       0.00937    0.12884    0.07     0.94
## black       -0.11189    0.18668   -0.60     0.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1468.3  on 1164  degrees of freedom
## Residual deviance: 1411.4  on 1161  degrees of freedom
## AIC: 1419
##
## Number of Fisher Scoring iterations: 4
```

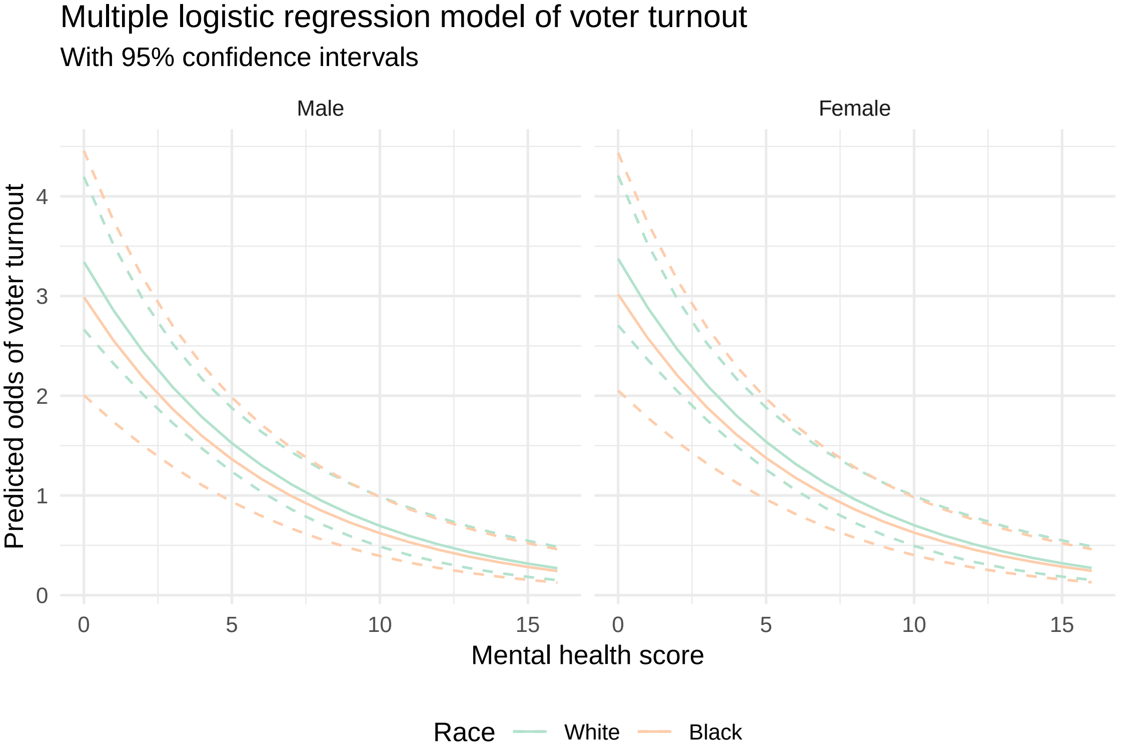

Multiple logistic regression model of voter turnout
With 95% confidence intervals
White
Black
Male
Female
Predicted log−odds of voter turnout
1
0
−1
−2
0
5
10
15
Mental health score


Multiple logistic regression model of voter turnout
With 95% confidence intervals
White
Black
Predicted log−odds of voter turnout
1
0
−1
−2
0
5
10
15
Mental health score
Gender
Male
Female


Multiple logistic regression model of voter turnout
With 95% confidence intervals
Male
Female
Predicted log−odds of voter turnout
1
0
−1
−2
0
5
10
15
Mental health score
Race
White
Black

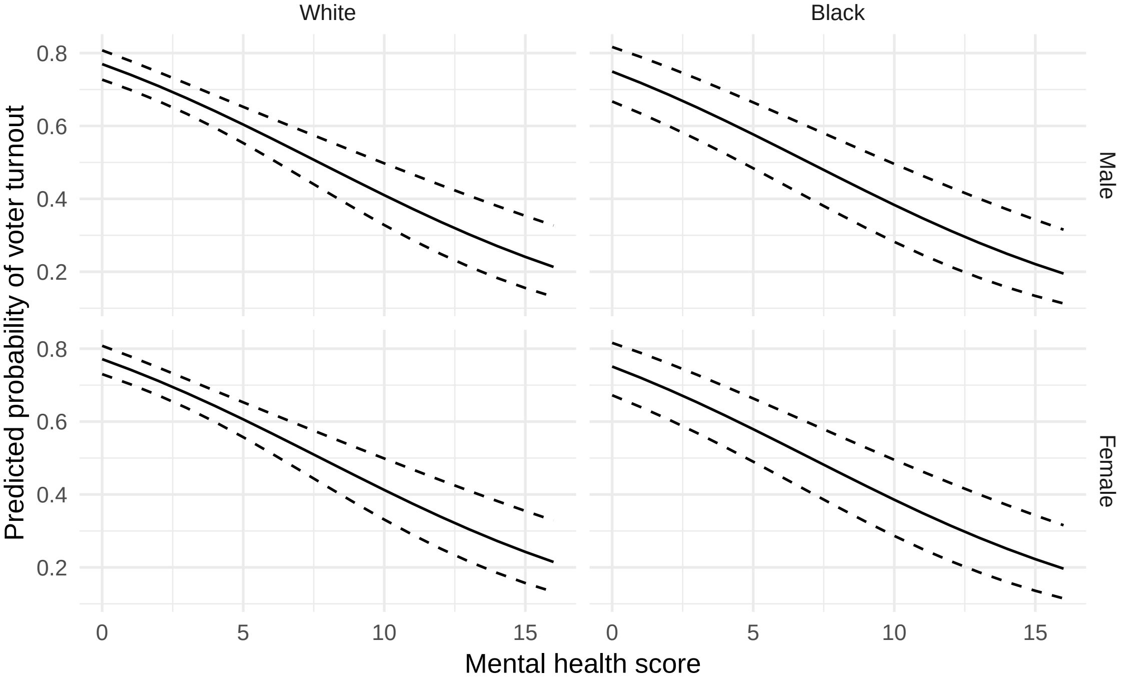

Multiple logistic regression model of voter turnout
With 95% confidence intervals
White
Black
Male
Female
Predicted odds of voter turnout
4
3
2
1
0
0
5
10
15
Mental health score


Multiple logistic regression model of voter turnout
With 95% confidence intervals
White
Black
Predicted odds of voter turnout
0
1
2
3
4
0
5
10
15
0
5
10
15
Mental health score
Gender
Male
Female


Multiple logistic regression model of voter turnout
With 95% confidence intervals
Male
Female
Predicted odds of voter turnout
0
1
2
3
4
0
5
10
15
0
5
10
15
Mental health score
Race
White
Black

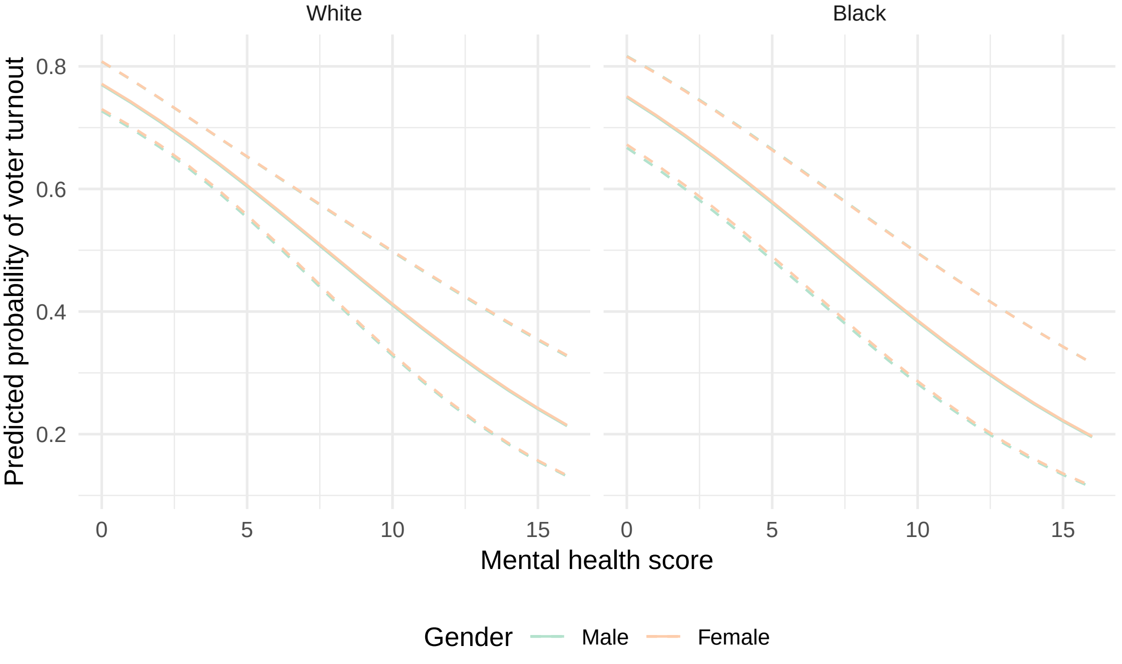
Multiple logistic regression model of voter turnout

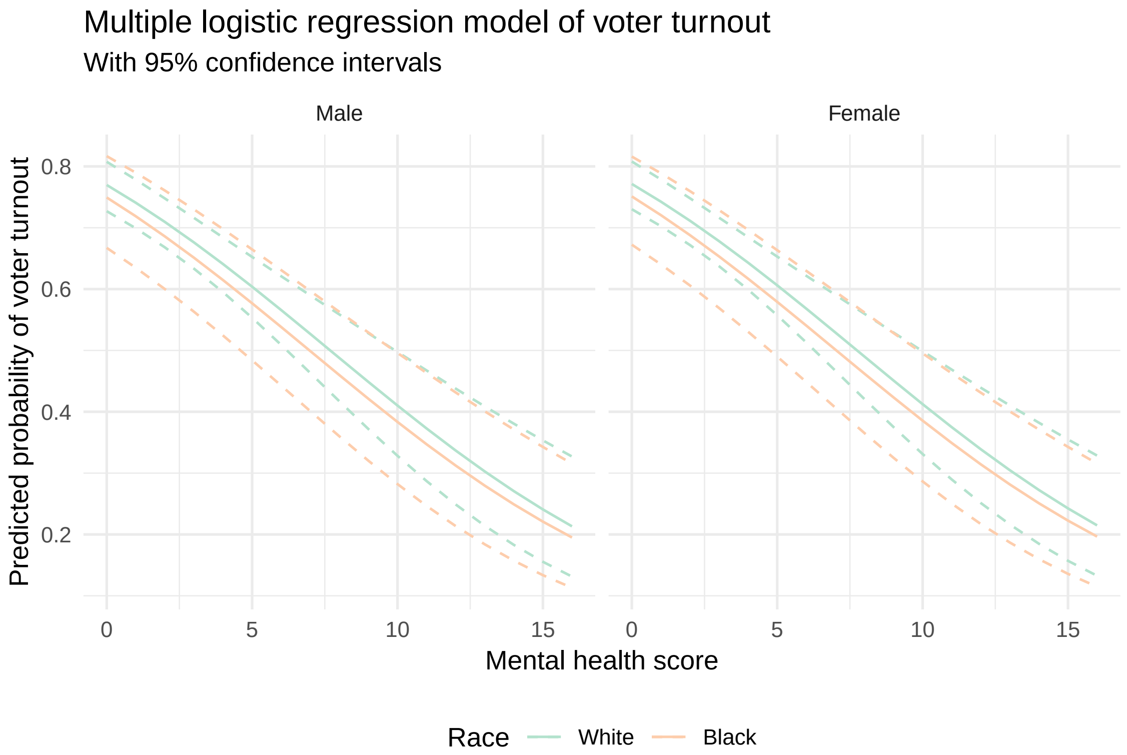
With 95% confidence intervals



Multiple logistic regression model of voter turnout

With 95% confidence intervals





Things students should do:

- Accurately interpret the resulting coefficients and standard errors. Don't confuse log-odds with odds with probabilities.
- Discuss the parameters in terms of substantive and statistical signficance **clearly** and draw a conclusion. Does the variable have a meaningful relationship with voter turnout or not?
- They should at some point address the relationship in terms of odds or probabilities. Log-odds are not intuitive at all.