Simple Logistic Regression

SDS 291

4/6/2020

We have data from each state (n=50) on their average income, education (% high school, % college, and % advanced degrees completed), political leaning from a 2015 Gallup poll and whether President Trump won that state (1=Win) or not (0=Did not Win) in the 2016 election.

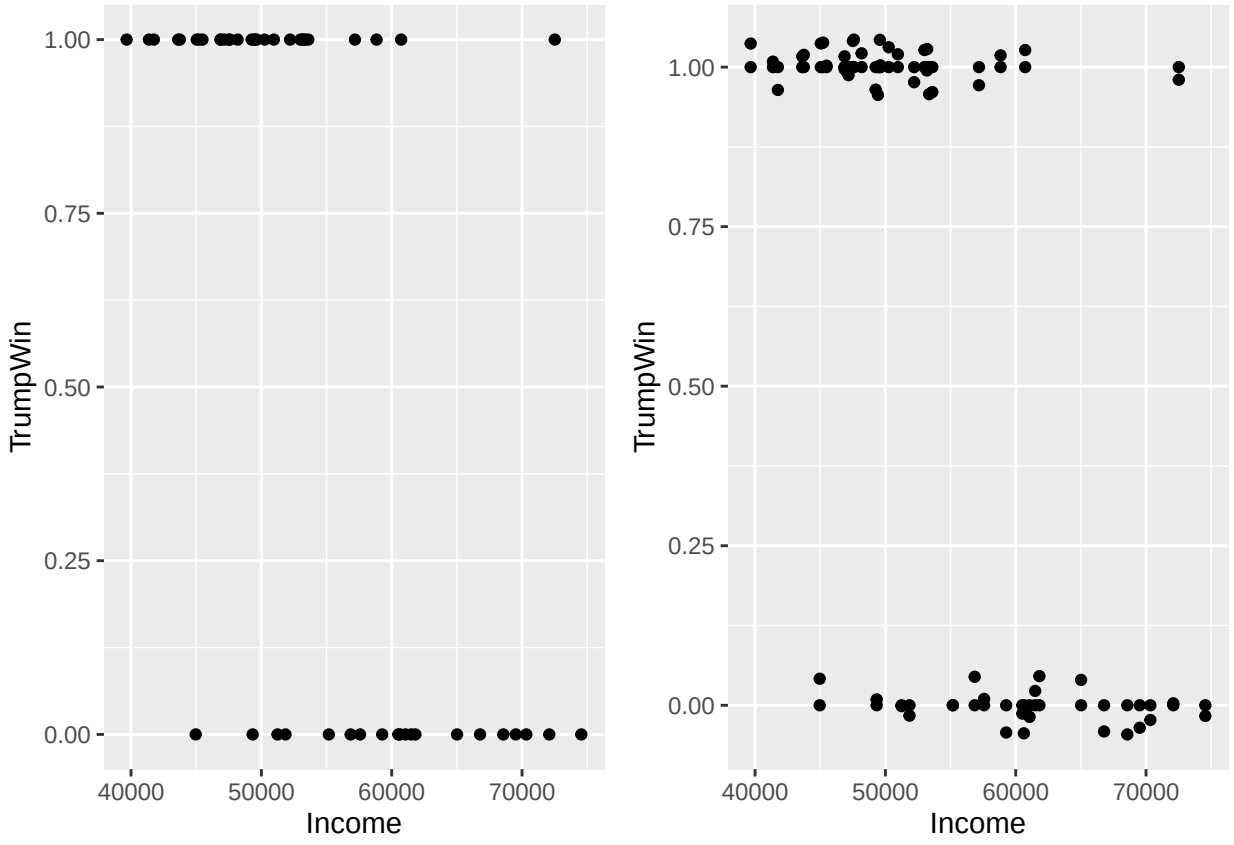
```
library(Stat2Data)
data("Election16")
```

Income and Election Outcome

Plots

Below are two plots exploring the relationship between income and President Trump winning that state. They are depicting the same pattern; the right "jitters" the data to spread the points out. (Note for the right plot, there aren't *actually* values >1 and <0, that's just a function of the actual data being spread out).

```
qplot(y=TrumpWin, x=Income, data=Election16)
qplot(y=TrumpWin, x=Income, data=Election16) + geom_jitter(width = 0.1, height=0.05)
```



1. Which is the easier graph to understand? Why?

2. What do you conclude from the plot about the relationship between income and the 2016 election results?

Logistic Model

Let's fit a logistic regression model to these data: $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1$

```
m0<-glm(TrumpWin~Income, data=Election16, family="binomial")
summary(m0)
```

```
##
## Call:
## glm(formula = TrumpWin ~ Income, family = "binomial", data = Election16)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2049  -0.7510   0.4074   0.6566   2.5000
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.118e+01  3.076e+00   3.635 0.000277 ***
## Income      -1.967e-04  5.582e-05  -3.523 0.000426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 67.301  on 49  degrees of freedom
## Residual deviance: 45.923  on 48  degrees of freedom
## AIC: 49.923
##
## Number of Fisher Scoring iterations: 5
```

We can write this fitted model as $log(odds) = 11.18 + -0.0001967 Income$

Let's use Income in $1,000s to make the interpretation a little easier. Then we re-fit a logistic regression model.

```
Election16<-Election16 %>% mutate(Income1000s = Income/1000)
m1<-glm(TrumpWin~Income1000s, data=Election16, family="binomial")
summary(m1)
```

```
##
## Call:
## glm(formula = TrumpWin ~ Income1000s, family = "binomial", data = Election16)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2049  -0.7510   0.4074   0.6566   2.5000
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.18186    3.07576   3.635 0.000277 ***
## Income1000s -0.19668    0.05582  -3.523 0.000426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 67.301  on 49  degrees of freedom
## Residual deviance: 45.923  on 48  degrees of freedom
## AIC: 49.923
##
## Number of Fisher Scoring iterations: 5
```

3. Write the fitted regression model equation using the output above.

4. What is the direction and magnitude of the relationship between the average income and whether Pres. Trump won that state?

5. Calculate the log(odds) (the book calls this the Empirical Logit), the odds, and the probability of President Trump winning for each of the following income levels. As a reminder, you can calculate each from the same output.

Log(odds):

$$log(odds) = \beta_0 + \beta_1 X_1$$

Odds:

$$Odds = e^{\beta_0 + \beta_1 X_1}$$

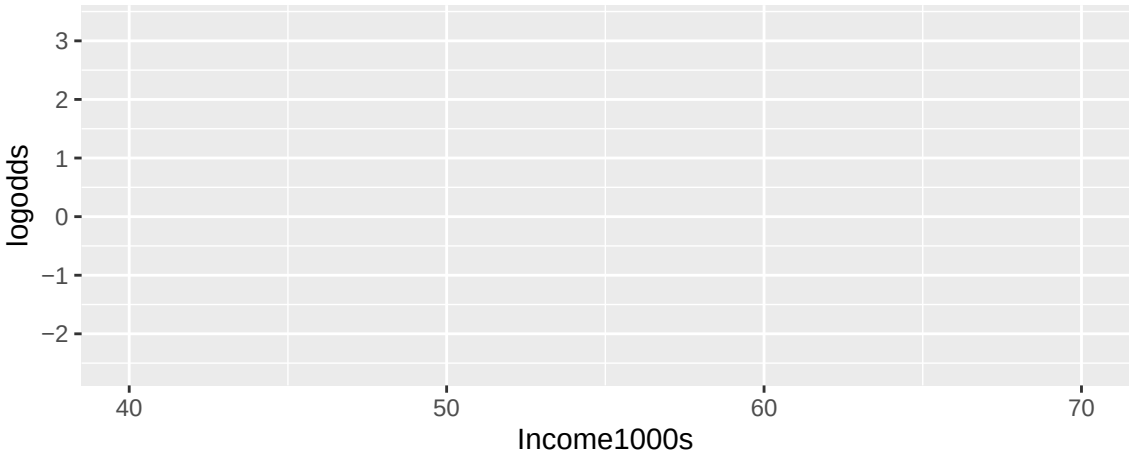
Probability:

$$\pi = \frac{odds}{1 + odds} = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

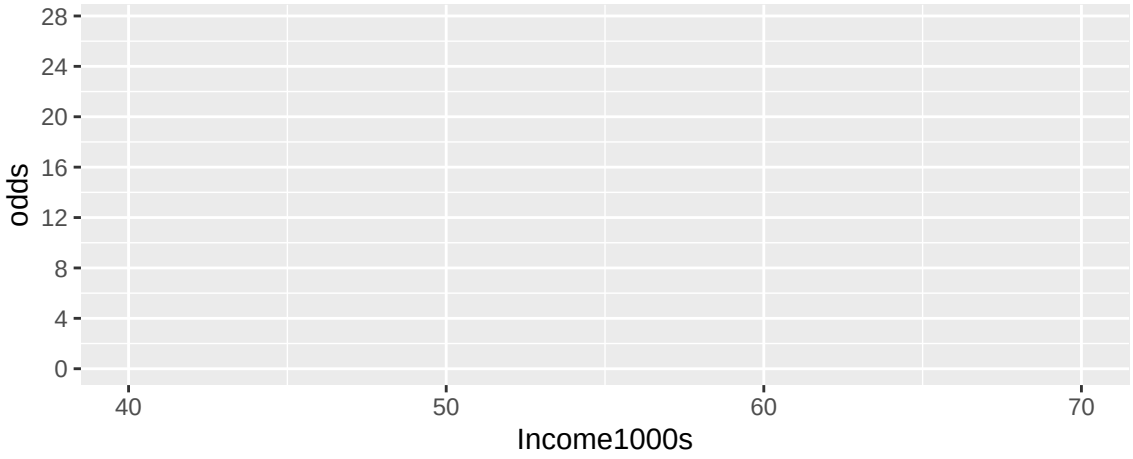
Income	$40,000	$50,000	$51,000	$55,000	$60,000	$61,000	$70,000
Log(odds)							
Odds							
Probability							

6. Plot the values on each of the three plots below.

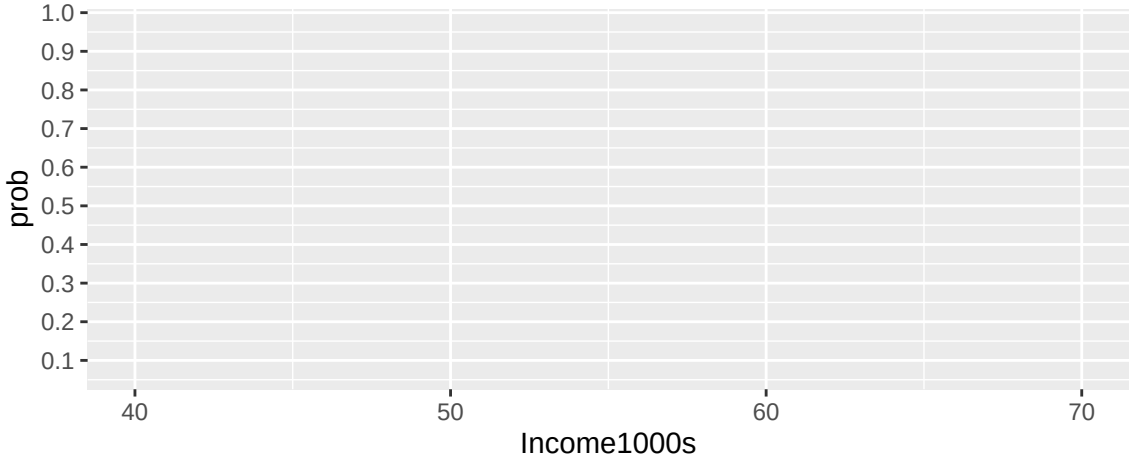
6a. Log(odds)



6b. Odds



6c. Probability



7. What is the ratio of odds for President Trump winning a state?

7a. Calculate the ratio the odds of a (theoretical) state with $51,000 average income to a state with $50,000 average income.

7b. Calculate the ratio the odds of a (theoretical) state with $61,000 average income to a state with $60,000 average income.

7c. Calculate the OR from the model ($OR = e^{\beta_1}$). Interpret the odds ratio in a sentence.

7d. Did you get the same values from each approach (7a-7c)? Why or why not?

8. Specify your hypotheses and conduct a test of whether the relationship between average income and President Trump winning a state is statistically signficant at the $\alpha = 0.05$ level.

9. Calculate the 95% Confidence Interval for the odds ratio of each additional $1,000 of average income and of Pres Trump winning that state. $t^* = 1.96$

Extra Practice

Create a binary variable of whether that state had above or below the national average rate of bachelors degree holders (35.6%) and repeat the steps above in `R`.

(*Hint*: Remember how to create a binary variable? See the IPUMS in-class exercise for examples of when you've done this before)