hmw6stat632

Jessica Grover

4/17/2022

# Reading the data

county\_votes16 <- read.csv("county\_votes16.csv")  
head(county\_votes16)

## state county clinton\_pctvotes trump\_pctvotes obama\_pctvotes pct\_pop65  
## 1 AL Autauga County 23.96 73.44 26.58 13.8  
## 2 AL Baldwin County 19.57 77.35 21.57 18.7  
## 3 AL Barbour County 46.66 52.27 51.25 16.5  
## 4 AL Bibb County 21.42 76.97 26.22 14.8  
## 5 AL Blount County 8.47 89.85 12.35 17.0  
## 6 AL Bullock County 75.09 24.23 76.31 14.9  
## pct\_black pct\_white pct\_hispanic pct\_asian highschool bachelors income  
## 1 18.7 77.9 2.7 1.1 85.6 20.9 53.682  
## 2 9.6 87.1 4.6 0.9 89.1 27.7 50.221  
## 3 47.6 50.2 4.5 0.5 73.7 13.4 32.911  
## 4 22.1 76.3 2.1 0.2 77.5 12.1 36.447  
## 5 1.8 96.0 8.7 0.3 77.0 12.1 44.145  
## 6 70.1 26.9 7.5 0.3 67.8 12.5 32.033  
## trump\_win  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 0

# Problem 1.

# a. Fit a simple logistic regression model with trump win as the binary response variable, and obama pctvotes as the predictor. Use summary() to print the results, and write down the equation for the estimated logistic regression model. Use this logistic regression model to answer the remaining questions.

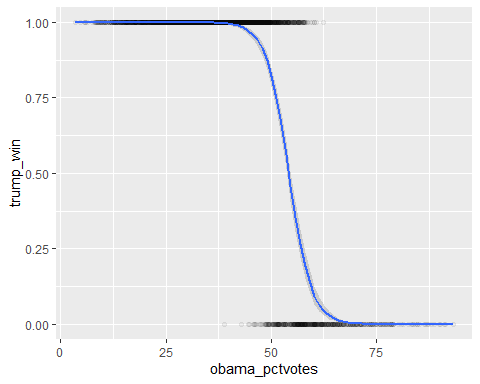
model1 <- glm(trump\_win ~ obama\_pctvotes, data = county\_votes16, family = "binomial")  
summary(model1)

##   
## Call:  
## glm(formula = trump\_win ~ obama\_pctvotes, family = "binomial",   
## data = county\_votes16)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3777 0.0025 0.0206 0.1159 2.4832   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 20.12971 1.04450 19.27 <2e-16 \*\*\*  
## obama\_pctvotes -0.37149 0.01971 -18.85 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2703.37 on 3111 degrees of freedom  
## Residual deviance: 736.42 on 3110 degrees of freedom  
## AIC: 740.42  
##   
## Number of Fisher Scoring iterations: 8

# b. ke a scatter plot of the data (i.e., plot the observed zeros and ones on the y-axis and obama pctvotes on the x-axis) and superimpose the fitted logistic curve for the estimated probability of Trump winning. Use ggplot2 to make the plot.

library(ggplot2)  
ggplot(county\_votes16, aes(x= obama\_pctvotes, y= trump\_win)) +  
geom\_point(alpha = 0.05) + geom\_smooth(method = "glm", method.args = "binomial")

## `geom\_smooth()` using formula 'y ~ x'

 # c. Use the logistic regression model to estimate the probability of Trump winning in a county with obama pctvotes = 40, 50, and 60.

votes <- data.frame(obama\_pctvotes = c(40,50,60))  
predict(model1, newdata = votes, type = "response")

## 1 2 3   
## 0.9948835 0.8256735 0.1034357

The probability of trump winning is 99.49% for 40% of obama votes. Simillarly, for 50% obama votesthe probablity of Trump winning is 82.56% and for 60% obama votes the probablity of Trump winning is 10.34%.

# d. Provide an interpretation of the estimated coefficient βˆ1 for obama pctvotes.

The estimated coefficient of Beta1Hat for obama\_pctvotes denotes that an increase in about one percent of votes that were casted for Obama in the year 2012 is realated to a decrease of about 0.37149 in the lod-odd of the probablity of Trump winning in 2016 polls.

# Problem 2.

# a. Fit a multiple logistic regression model with trump win as the response, and the following 8 demographic variables as predictors: pct pop65, pct black, pct white, pct hispanic, pct asian, highschool, bachelors, and income. Use summary() to print the results.

model2 <- glm(trump\_win ~ pct\_pop65 + pct\_black + pct\_white + pct\_hispanic + pct\_asian + highschool + bachelors+income+trump\_win,data=county\_votes16,family = "binomial")

## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the  
## right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 9 in  
## model.matrix: no columns are assigned

summary(model2)

##   
## Call:  
## glm(formula = trump\_win ~ pct\_pop65 + pct\_black + pct\_white +   
## pct\_hispanic + pct\_asian + highschool + bachelors + income +   
## trump\_win, family = "binomial", data = county\_votes16)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2155 0.0648 0.1350 0.3170 2.9283   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.760459 1.721412 1.604 0.108802   
## pct\_pop65 -0.020445 0.017910 -1.142 0.253632   
## pct\_black -0.035455 0.007739 -4.581 4.63e-06 \*\*\*  
## pct\_white 0.084759 0.007873 10.765 < 2e-16 \*\*\*  
## pct\_hispanic -0.083716 0.007005 -11.952 < 2e-16 \*\*\*  
## pct\_asian -0.160999 0.046158 -3.488 0.000487 \*\*\*  
## highschool -0.042242 0.020994 -2.012 0.044204 \*   
## bachelors -0.193758 0.014444 -13.415 < 2e-16 \*\*\*  
## income 0.048985 0.008503 5.761 8.39e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2703.4 on 3111 degrees of freedom  
## Residual deviance: 1269.4 on 3103 degrees of freedom  
## AIC: 1287.4  
##   
## Number of Fisher Scoring iterations: 7

# b. Remove any predictors that are not significant from the model fit in (a).

model3 <- glm(trump\_win ~ pct\_black + pct\_white + pct\_hispanic + pct\_asian + highschool + bachelors + income+trump\_win,data=county\_votes16,family = "binomial")

## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the  
## right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 8 in  
## model.matrix: no columns are assigned

summary(model3)

##   
## Call:  
## glm(formula = trump\_win ~ pct\_black + pct\_white + pct\_hispanic +   
## pct\_asian + highschool + bachelors + income + trump\_win,   
## family = "binomial", data = county\_votes16)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2004 0.0653 0.1351 0.3205 2.9272   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.644244 1.716668 1.540 0.123479   
## pct\_black -0.036567 0.007694 -4.753 2.01e-06 \*\*\*  
## pct\_white 0.081996 0.007479 10.963 < 2e-16 \*\*\*  
## pct\_hispanic -0.082609 0.006918 -11.942 < 2e-16 \*\*\*  
## pct\_asian -0.152133 0.044877 -3.390 0.000699 \*\*\*  
## highschool -0.043707 0.020911 -2.090 0.036606 \*   
## bachelors -0.192417 0.014389 -13.373 < 2e-16 \*\*\*  
## income 0.050576 0.008376 6.039 1.56e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2703.4 on 3111 degrees of freedom  
## Residual deviance: 1270.7 on 3104 degrees of freedom  
## AIC: 1286.7  
##   
## Number of Fisher Scoring iterations: 7

We are removing pct\_pop65 as it is not significant.

# c. Provide an interpretation of the signs of the estimated coefficients.

There is a positive relationship with income and percentage of white people vote. Negative coefficients are negatively related to the response variable. null deviance- residual deviance and with the number of predictors as 6 we get the value of p less than 0.05 in which we can say that the model is significant with 95 percent confidence.