Homework 6

Winnie Lu

2022-04-12

```
county_votes16 <-read.csv("county_votes16.csv")
```

Exercise 1

a

- $log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x = 20.13 - 0.3715x$

```
glm1 <-glm(trump_win ~ obama_pctvotes, data = county_votes16, family = binomial)
summary(glm1)
```

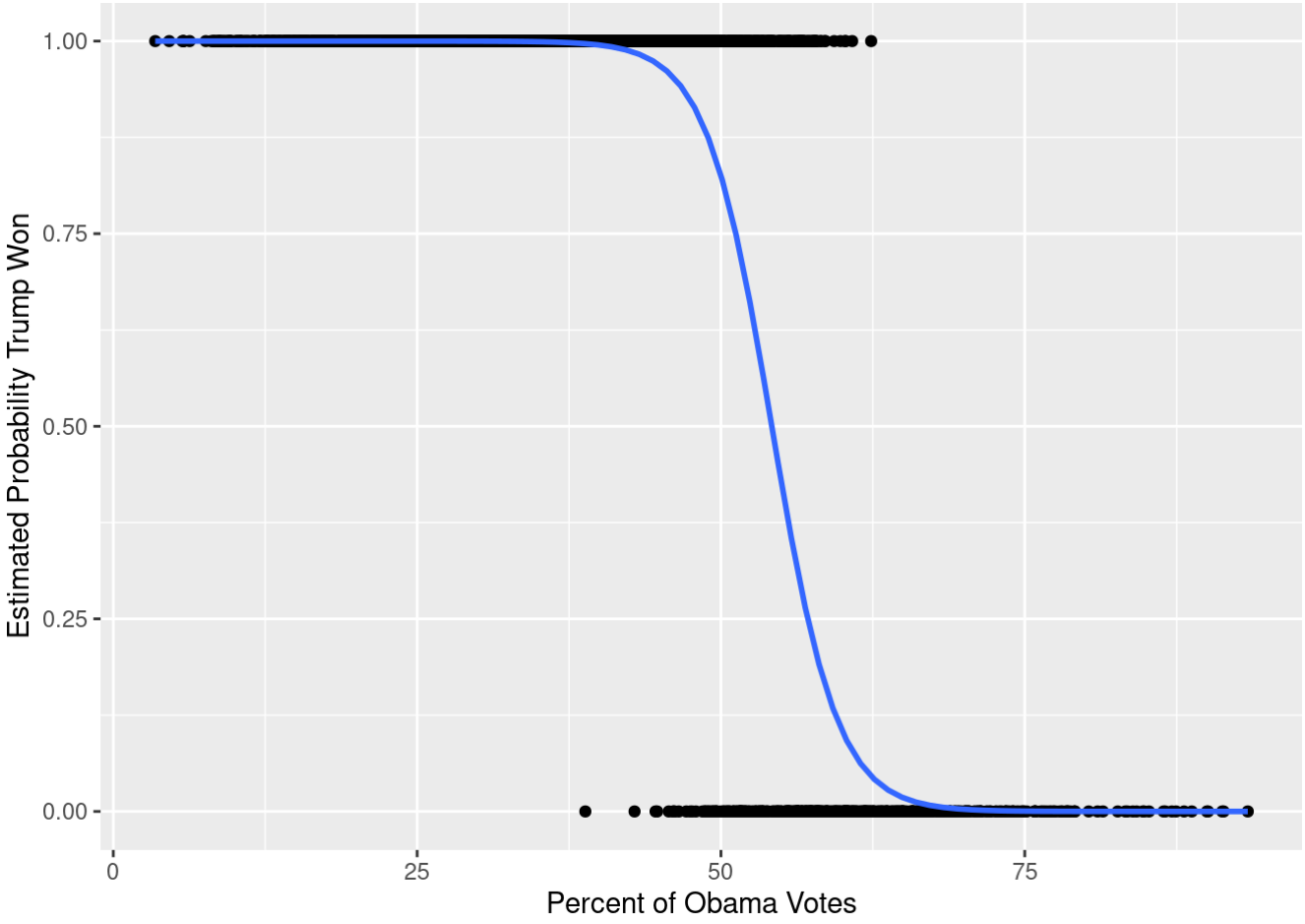
```
##
## Call:
## glm(formula = trump_win ~ obama_pctvotes, family = binomial,
##     data = county_votes16)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3777   0.0025   0.0206   0.1159   2.4832
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (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

- Plot showing the logistic probability Trump wins using percent of Obama votes as a predictor.

```
library(ggplot2)
ggplot(glm1, aes(obama_pctvotes, trump_win)) + geom_point() +
        geom_smooth(method="glm", method.args =list(family="binomial"), se=F) +
    xlab("Percent of Obama Votes") +
    ylab("Estimated Probability Trump Won")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



C

- Given that the percent of Obama vote is 40, the estimated probability that Trump wins is 99%. If the percent of Obama votes is 50, the estimated probability that Trump Wins is 82.57%. Lastly, if the the percent of votes cast for Obama is 60%, the estimated proability that Trump wins is 10.34%.

```
new_x1 <-data.frame(obama_pctvotes = 40)
new_x2 <-data.frame(obama_pctvotes = 50)
new_x3 <-data.frame(obama_pctvotes = 60)
predict(glm1, newdata=new_x1, type="response")
```

```
##         1
## 0.9948835
```

```
predict(glm1, newdata=new_x2, type="response")
```

```
##         1
## 0.8256735
```

```
predict(glm1, newdata=new_x3, type="response")
```

```
##         1
## 0.1034357
```

d

- In terms of the odds, if the percentage of obama votes in a particular county is less than 54.18%, then Trump wins, if it's more than 58.17% then Trump loses. Since $\hat{\beta}_1 < 0$, then increasing the percentage of Obama votes will be associated with decreasing the probability that Trump wins.

Exercise 2

a

```
glm2 <-glm(trump_win ~ pct_pop65 + pct_black + pct_white + pct_hispanic + pct_asian + hi
ghschool + bachelors + income, data=county_votes16, family=binomial)
summary(glm2)
```

```
##
## Call:
## glm(formula = trump_win ~ pct_pop65 + pct_black + pct_white +
##     pct_hispanic + pct_asian + highschool + bachelors + income,
##     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

- The predictors that are not significant in our model according to an $\alpha = 0.05$ are: pct_pop65, or the percent of votes over 65 years old. All other remaining predictors are significant. The third multiple logistic model (yielded with the step() function) shows that after removing pct_pop65, the remaining predictors remain significant. To confirm our results, we yield the AIC output for both models and confirm that our glm_sel model has the lower AIC;therefore, it is the better model.

```
glm2 <-glm(trump_win ~ pct_pop65 + pct_black + pct_white + pct_hispanic + pct_asian + hi
ghschool + bachelors + income, data=county_votes16, family=binomial)

glm_sel <-step(glm2)
```

```
## Start:  AIC=1287.42
## trump_win ~ pct_pop65 + pct_black + pct_white + pct_hispanic +
##     pct_asian + highschool + bachelors + income
##
##                Df Deviance    AIC
## - pct_pop65     1   1270.7 1286.7
## <none>              1269.4 1287.4
## - highschool    1   1273.5 1289.5
## - pct_asian     1   1283.6 1299.6
## - pct_black     1   1289.5 1305.5
## - income        1   1304.8 1320.8
## - pct_white     1   1373.3 1389.3
## - pct_hispanic  1   1438.4 1454.4
## - bachelors     1   1489.3 1505.3
##
## Step:  AIC=1286.71
## trump_win ~ pct_black + pct_white + pct_hispanic + pct_asian +
##     highschool + bachelors + income
##
##                Df Deviance    AIC
## <none>              1270.7 1286.7
## - highschool    1   1275.2 1289.2
## - pct_asian     1   1284.0 1298.0
## - pct_black     1   1292.2 1306.2
## - income        1   1310.0 1324.0
## - pct_white     1   1374.5 1388.5
## - pct_hispanic  1   1438.9 1452.9
## - bachelors     1   1489.4 1503.4
```

```
summary(glm_sel)
```

```
##
## Call:
## glm(formula = trump_win ~ pct_black + pct_white + pct_hispanic +
##     pct_asian + highschool + bachelors + income, 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
```

```
AIC(glm2, glm_sel)
```

```
##         df      AIC
## glm2     9 1287.424
## glm_sel  8 1286.710
```

C

- For our “final” model, we notice that with an increase in percent of black, hispanic, asian, highschool, and bachelors-degree voters, there is an associated decrease in the probability that Trumps wins. With an increase in Black voters by 1%, there is an associated decrease in the probability that Trump wins by 3.6%. With an increase in Hispanic voters by 1%, there is an associated decrease in the probability that Trump wins by 8.3%. With an increase in Asian voters by 1%, there is an associated decrease in the probability that Trump wins by 1.5%. With an increase in Highschool voters by 1%, there is an associated decrease in the probability that Trump wins by 4.4%. With an increase in Bachelors-degree voters by 1%, there is an associated decrease in the probability that Trump wins by 1.9%. In contrast, with an increase in White voters by 1%, there is an associated increase in the proability that Trumps wins by 8.2%. Lastly, with every $1000 increase in income, there is an associated increase in the proability that Trump wins by 5.1%.