Logistic Regression - Answers

SDS 291

4/6/2020

Contents

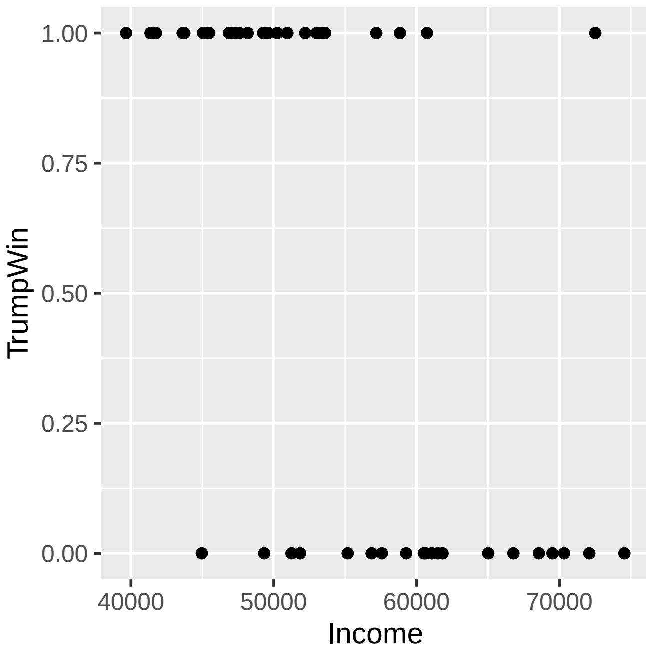
We have data from each state (n=50) on their average income, education (% high school, % college, and % advanced degrees completed), political leaning from a 2015 Gallup poll and whether President Trump won that state (1=Win) or not (0=Did not Win) in the 2016 election.

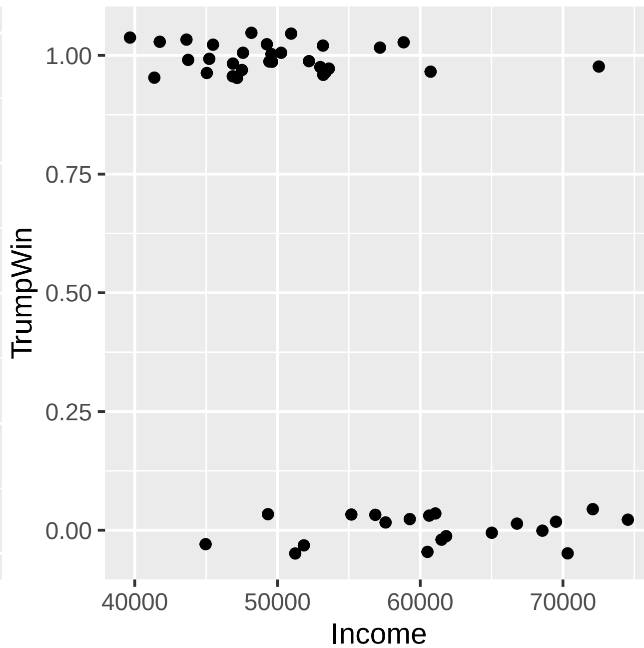
```
## 'data.frame':    50 obs. of  8 variables:
##  $ State   : Factor w/ 50 levels "Alabama ","Alaska ",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Abr     : Factor w/ 50 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 8 9 10 ...
##  $ Income  : int  43623 72515 50255 41371 61818 60629 70331 60509 47507 49620 ...
##  $ HS      : num  84.3 92.1 86 84.8 81.8 90.7 89.9 88.4 86.9 85.4 ...
##  $ BA      : num  23.5 28 27.5 21.1 31.4 38.1 37.6 30 27.3 28.8 ...
##  $ Adv     : num  8.7 10.1 10.2 7.5 11.6 14 16.6 12.2 9.8 10.7 ...
##  $ Dem.Rep : int  -17 -17 -1 -7 16 -1 11 6 1 -4 ...
##  $ TrumpWin: int  1 1 1 1 0 0 0 0 1 1 ...
```

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 second "jitters" the data.





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

The right plot, with the jittered data makes the patterns more apparent since the points aren't overlaying.

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

The plot depicts a negative relationship between income and the 2016 election results. It is hard to determine the magnitude, but it seems sort of moderately to weak, since there are a number of states with the same average income that President Trump won and didn't win.

Logistic Model

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

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)
summary(m1)
```

```
##
## Call:
## glm(formula = TrumpWin ~ Income1000s, data = Election16)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.91533  -0.31986   0.08508   0.24138   1.01398
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.431908   0.348034   6.988 7.68e-09 ***
## Income1000s -0.033729   0.006324  -5.333 2.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 0.1569807)
##
##     Null deviance: 12.0000  on 49  degrees of freedom
## Residual deviance:  7.5351  on 48  degrees of freedom
## AIC: 53.271
##
## Number of Fisher Scoring iterations: 2
```

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

$log(odds) = 2.432 + -0.03373 Income1000s$

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

The association is negative and slight/shallow slope.

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:

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

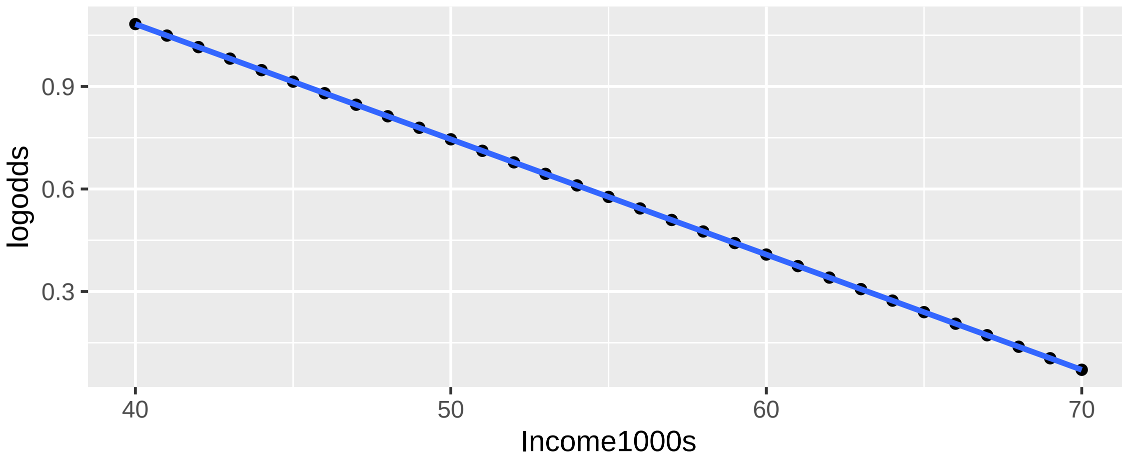
```
Income1000s<-c(40:70)
logodds<-coef(m1)[1]+ (coef(m1)[2]*Income1000s)
m1_data<-as.data.frame(cbind(Income1000s,logodds))
m1_data <- m1_data %>%
  mutate(odds=exp(logodds),
         prob = odds/(1+odds))
```

Income	$40,000	$50,000	$51,000	$55,000	$60,000	$61,000	$70,000
Log(odds)	1.0827307	0.7454365	0.7117071	0.5767894	0.4081423	0.3744129	0.0708481
Odds	2.9527317	2.1073612	2.0374665	1.7803134	1.5040212	1.4541375	1.0734182
Probability	0.7470104	0.6781835	0.6707783	0.6403283	0.6006424	0.5925249	0.5177046

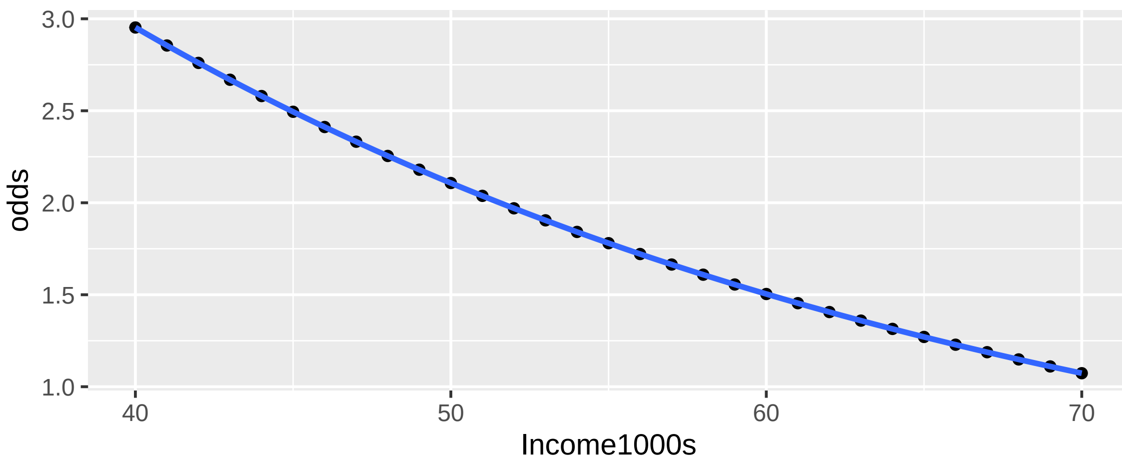
```
## Observations: 31
## Variables: 4
## $ Income1000s <dbl> 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53,...
## $ logodds     <dbl> 1.0827307, 1.0490013, 1.0152719, 0.9815425, 0.9478131, ...
## $ odds        <dbl> 2.952732, 2.854799, 2.760114, 2.668569, 2.580061, 2.494...
## $ prob        <dbl> 0.7470104, 0.7405831, 0.7340506, 0.7274142, 0.7206752, ...
```

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

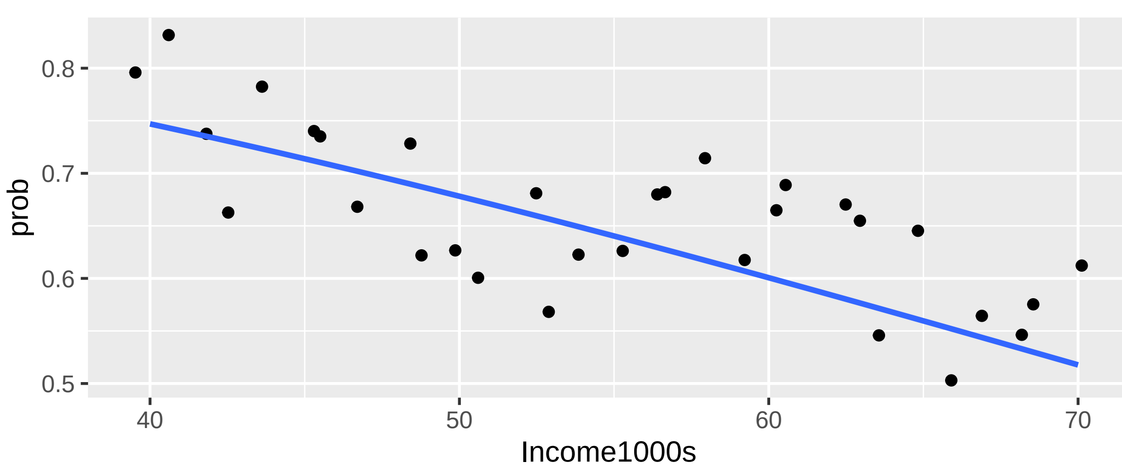
Log(odds)

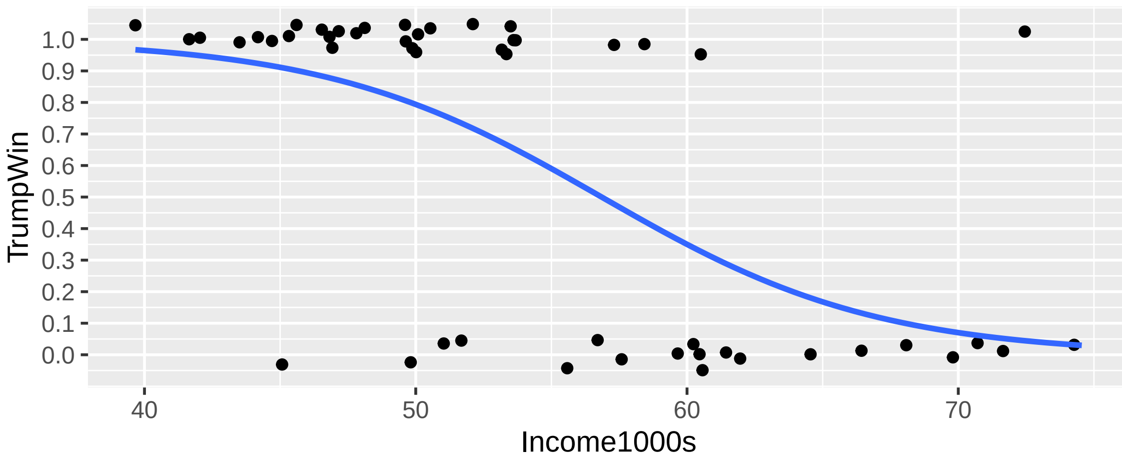


Odds



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.

```
Odds51<-2.037466
Odds50<-2.107361
OR5051<-Odds51/Odds50
OR5051
```

```
## [1] 0.9668329
```

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

```
Odds61<-1.454137
Odds60<-1.504021
OR6061<-Odds61/Odds60
OR6061
```

```
## [1] 0.9668329
```

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

```
exp(coef(m1)[2])
```

```
## Income1000s
##   0.9668331
```

Each additional \$1,000 average income was associated with 0.9668 times the odds of Trump winning in a state. States with \$1,000 higher average income had a lower odds of Trump winning than a state with \$1,000 lower average income.

7d. Did you get the same values from each approach? Why or why not?

Yes. Since we fit the logistic regression model as a linear model on the log(odds) scale, the slope of the line, or the Odds Ratio, is constant, just as a slope was in linear regression.

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.

$H_0 : \beta_1 = 0$

$H_0 : \beta_1 \neq 0$

We reject the null hypothesis since the test statistic for the relationship between income and Trump winning was below the critical value (-5.333<-1.96) and the associated p-value is below the 0.05 threshold (p<0.001).

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. $t^* = 1.96$

```
lci<-exp(-0.033729 - (1.96*0.006324))
uci<-exp(-0.033729 + (1.96*0.006324))
CIs<-cbind(lci,uci)
CIs
```

```
##            lci      uci
## [1,] 0.9549235 0.978892
```

Or get `R` to do it for you:

```
exp(confint(m1))
```

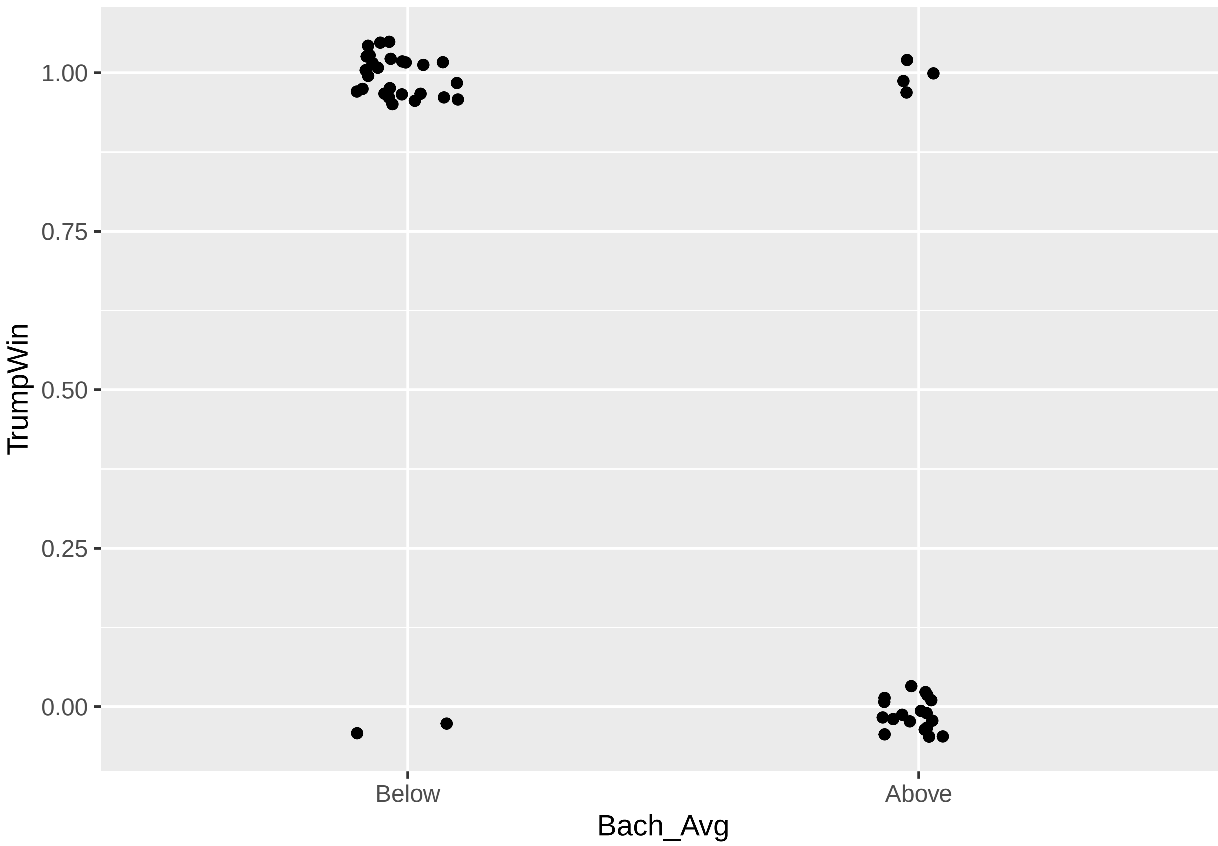
```
## Waiting for profiling to be done...
```

```
##                 2.5 %     97.5 %
## (Intercept) 5.7532935 22.5118658
## Income1000s 0.9549224  0.9788923
```

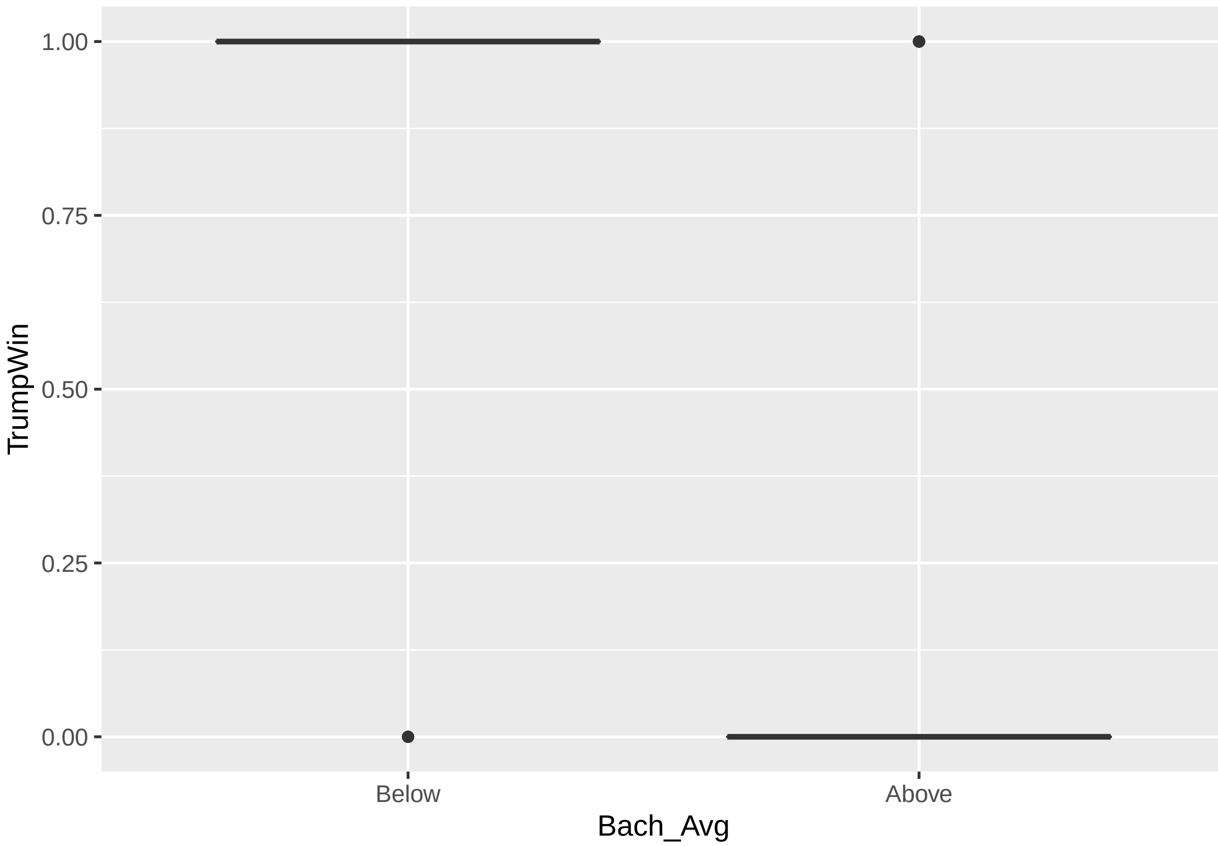
10. Create a binary variable of whether that state had above or below the national average rate of bachelor's degree holders (29.1%) and repeat the steps above in `R`. (A note, these are the averages from these data, rather than from an external source that quoted a higher value. You might want to dichomoize at the median rather than the mean to avoid some very high/low BA degree states skewing the data, but nonethelesss, this is one approach.)

```
Election16<-Election16 %>%
  mutate(Bach_Avg=as.factor(if_else(BA>=29,"Above","Below")))
Election16$Bach_Avg <- relevel(Election16$Bach_Avg, ref="Below")

Election16 %>%
  ggplot(aes(y=TrumpWin, x=Bach_Avg)) + geom_jitter(width=0.1, height=0.05)
```




```
#another way to look at this is with a boxplot -- it will give you a solid line to indicate where most
Election16 %>%
  ggplot(aes(y=TrumpWin, x=Bach_Avg)) + geom_boxplot()
```



We see from the visual that being in a state that has higher than average rates of bachelor's degree holders is negatively associated with Trump winning those states.

```
library(mosaic)
tally(TrumpWin~Bach_Avg, data=Election16)
```

```
##          Bach_Avg
## TrumpWin Below Above
##        0     2    18
##        1    26     4
```

```
tally(TrumpWin~Bach_Avg, margins=TRUE, format = "proportion", data=Election16)
```

```
##          Bach_Avg
## TrumpWin      Below      Above
##    0     0.07142857 0.81818182
##    1     0.92857143 0.18181818
##    Total 1.00000000 1.00000000
```

```
#
library(gmodels)
with(Election16, CrossTable(TrumpWin, Bach_Avg,
prop.r=FALSE, prop.chisq=FALSE, prop.t=FALSE))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
```

```
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  50
##
##
##              | Bach_Avg
##     TrumpWin |     Below |     Above | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |         2 |        18 |        20 |
##              |     0.071 |     0.818 |           |
## -------------|-----------|-----------|-----------|
##            1 |        26 |         4 |        30 |
##              |     0.929 |     0.182 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        28 |        22 |        50 |
##              |     0.560 |     0.440 |           |
## -------------|-----------|-----------|-----------|
##
##
```

Models

```
m2_2016<-glm(TrumpWin~Bach_Avg, data=Election16)
summary(m2_2016)
```

```
##
## Call:
## glm(formula = TrumpWin ~ Bach_Avg, data = Election16)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.92857  -0.18182   0.07143   0.07143   0.81818
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.92857    0.06178  15.030  < 2e-16 ***
## Bach_AvgAbove -0.74675    0.09314  -8.018 2.07e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1068723)
##
##     Null deviance: 12.0000  on 49  degrees of freedom
## Residual deviance:  5.1299  on 48  degrees of freedom
## AIC: 34.047
##
## Number of Fisher Scoring iterations: 2
```

```
#This code combines the odds ratios and the 95% CIs into three columns (OR, lower CI, Upper CI) for eac
results_2016<-exp(cbind(coef(m2_2016), confint(m2_2016)))
```

```
## Waiting for profiling to be done...
```

```r
results_2016
```

```
##                              2.5 %    97.5 %
## (Intercept)   2.5308910 2.2422577 2.8566785
## Bach_AvgAbove 0.4739027 0.3948298 0.5688116
```

A state with more than the national average of bachelor's degree holders (i.e., a more highly educated state) has 0.47 times lower odds of Trump winning the state than a state with less than the national average of bachelor degree holders. We are 95% confident that the true relationship between having above average amount of bachelor's degree holders, compared to below average, and odds Trump winning is between 0.39 and 0.57. We can reject the null hypothesis ($H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$) and conclude that the association between a state having an above average number of bachelor degree holders had significantly lower odds of Trump winning than a state with below average bachelor's degree holders, as evidenced by the large test statstic (-8.018 < -1.96) and the small p-value (2.07e-10, which is below our threshold of 0.05).

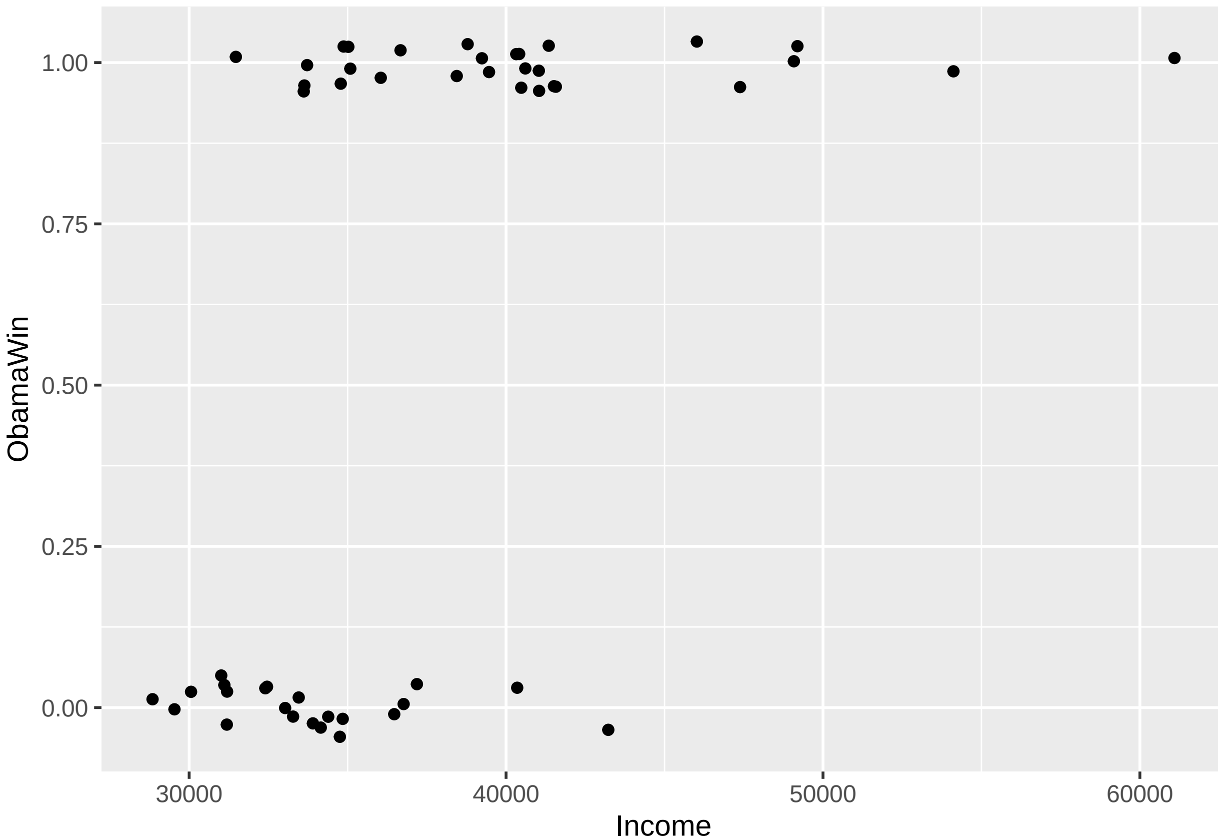
Recreating the above with the Elections 2008 data

The `Election16` data is in version 2.0 of the Stat2Data package, which isn't yet on the server in the Stat2Data package. You could practice with the `Election08` data, with the same variables based on the 2008 election.

```r
data("Election08")
Election08<-Election08 %>%
  mutate(Bach_Avg=as.factor(if_else(BA>=27.14,"Above","Below")),
    Income1000s = Income/1000)
Election08$Bach_Avg <- relevel(Election08$Bach_Avg, ref="Below")
```

#Obama Winning and Income

```r
Election08 %>%
  ggplot(aes(y=ObamaWin, x=Income)) + geom_jitter(width = 0.1, height=0.05)
```




```r
m1_2008<-glm(ObamaWin~Income1000s, data=Election08)
summary(m1_2008)
```

```
##
## Call:
## glm(formula = ObamaWin ~ Income1000s, data = Election08)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.79853  -0.38390  -0.03961   0.36100   0.68528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.981000   0.361775  -2.712  0.00921 **
## Income1000s  0.041168   0.009477   4.344 7.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1843239)
##
##     Null deviance: 12.5098  on 50  degrees of freedom
## Residual deviance:  9.0319  on 49  degrees of freedom
## AIC: 62.447
##
## Number of Fisher Scoring iterations: 2
```

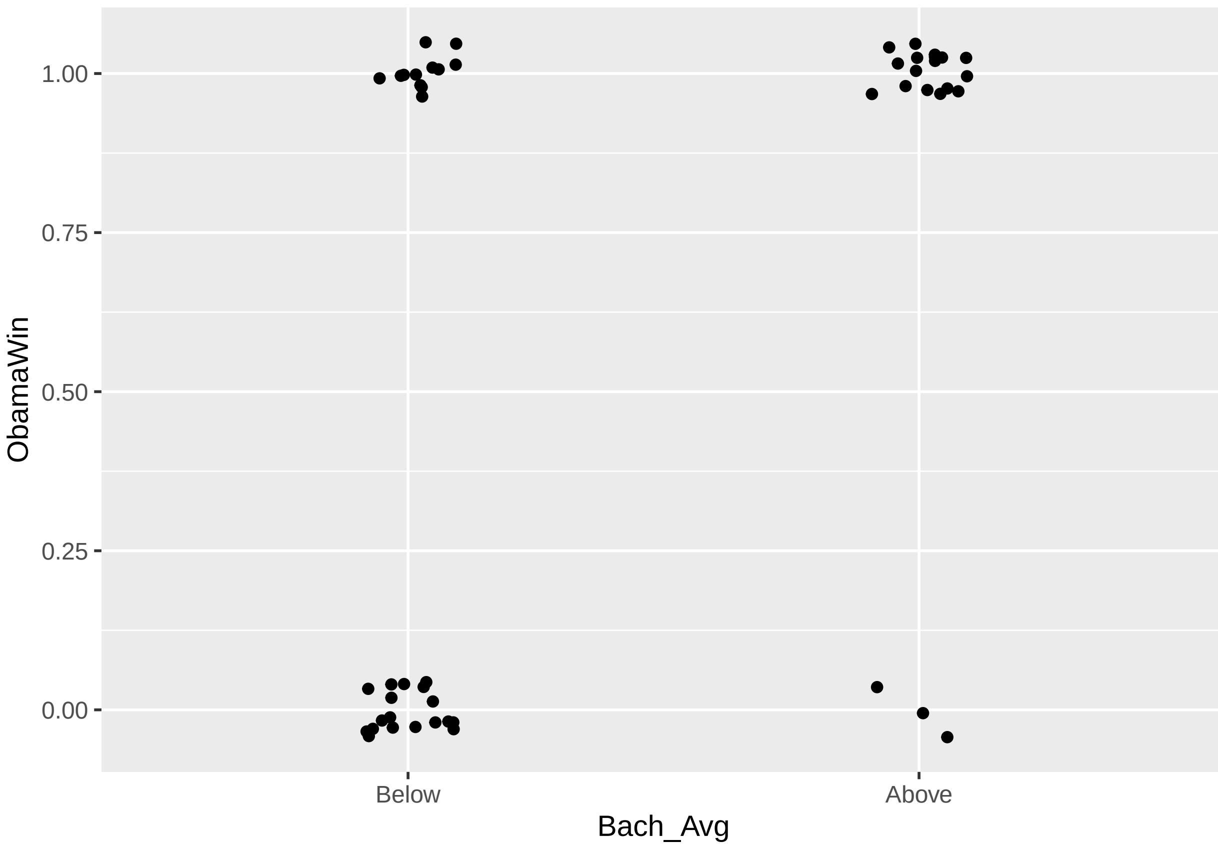
```r
#This code combines the odds ratios and the 95% CIs into three columns (OR, lower CI, Upper CI) for eac
results1_2008<-exp(cbind(coef(m1_2008), confint(m1_2008)))
```

```r
results1_2008
```

```
##                   2.5 %    97.5 %
## (Intercept) 0.3749362 0.1845076 0.7619044
## Income1000s 1.0420271 1.0228497 1.0615641
```

#Obama Winning and Bachelor's Degrees

```r
Election08 %>%
  ggplot(aes(y=ObamaWin, x=Bach_Avg)) + geom_jitter(width = 0.1, height=0.05)
```




```r
m2_2008<-glm(ObamaWin~Bach_Avg, data=Election08)
summary(m2_2008)
```

```
##
## Call:
## glm(formula = ObamaWin ~ Bach_Avg, data = Election08)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8500  -0.3871   0.1500   0.1500   0.6129
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.38710    0.08075   4.794 1.56e-05 ***
## Bach_AvgAbove 0.46290    0.12895   3.590 0.000764 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 0.2021396)
##
##     Null deviance: 12.5098  on 50  degrees of freedom
## Residual deviance:  9.9048  on 49  degrees of freedom
## AIC: 67.153
##
## Number of Fisher Scoring iterations: 2
```

```
#This code combines the odds ratios and the 95% CIs into three columns (OR, lower CI, Upper CI) for eac
results2_2008<-exp(cbind(coef(m2_2008), confint(m2_2008)))
results2_2008
```

```
##                          2.5 %   97.5 %
## (Intercept)   1.472699 1.257127 1.725237
## Bach_AvgAbove 1.588680 1.233887 2.045489
```