|  |  |  |  |
| --- | --- | --- | --- |
| **PM592: Regression Analysis for Health Data Science** |  |  |  |
| **Lab 8 – Examining Associations among Binary Variables**  **Data Needed:** *okc\_profiles\_cleaned.csv* | | | |

**Outline**

* Contingency Tables
* Odds Ratio
* Model Assessment
* Predictions

1. **Contingency Tables**
   1. **Table**
      1. A simple table can be created using the table() function. Suppose we want to examine the relationship between gender and being coupled (married or “seeing someone” vs. single).

> okc %>%

+ with(., table(male, coupled))

coupled

male 0 1

0 22979 1138

1 34590 1236

* + 1. We can use prop.table() on a table object to produce the proportions, optionally specifying row percentages (margin = 1) or column percentages (margin = 2)

> okc %>%

+ with(., table(male, coupled)) %>%

+ prop.table(margin = 1)

coupled

male 0 1

0 0.95281337 0.04718663

1 0.96549992 0.03450008

* 1. Xtabs
     1. xtabs(), or cross-tabs, is another way to make a table
     2. You can perform follow-up functions, like prop.table() or chisq.test(), using this method as well.
     3. xtabs can be used to make a table by specifying frequencies of each covariate pattern (see lecture).
  2. CrossTable
     1. A full complement of N, row, column, and overall percentages can be obtained by CrossTable() in gmodels.

> okc %>%

+ with(., table(male, coupled)) %>%

+ gmodels::CrossTable()

Cell Contents

|-------------------------|

| N |

| Chi-square contribution |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 59943

| coupled

male | 0 | 1 | Row Total |

-------------|-----------|-----------|-----------|

0 | 22979 | 1138 | 24117 |

| 1.444 | 35.010 | |

| 0.953 | 0.047 | 0.402 |

| 0.399 | 0.479 | |

| 0.383 | 0.019 | |

-------------|-----------|-----------|-----------|

1 | 34590 | 1236 | 35826 |

| 0.972 | 23.567 | |

| 0.965 | 0.035 | 0.598 |

| 0.601 | 0.521 | |

| 0.577 | 0.021 | |

-------------|-----------|-----------|-----------|

Column Total | 57569 | 2374 | 59943 |

| 0.960 | 0.040 | |

-------------|-----------|-----------|-----------|

* 1. The Chi-Square test
     1. These methods all include the ability to conduct a chi-square test of association between the X and Y variable.

> okc %>%

+ with(., table(male, coupled)) %>%

+ gmodels::CrossTable(chisq = T)

Cell Contents

|-------------------------|

| N |

| Chi-square contribution |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 59943

| coupled

male | 0 | 1 | Row Total |

-------------|-----------|-----------|-----------|

0 | 22979 | 1138 | 24117 |

| 1.444 | 35.010 | |

| 0.953 | 0.047 | 0.402 |

| 0.399 | 0.479 | |

| 0.383 | 0.019 | |

-------------|-----------|-----------|-----------|

1 | 34590 | 1236 | 35826 |

| 0.972 | 23.567 | |

| 0.965 | 0.035 | 0.598 |

| 0.601 | 0.521 | |

| 0.577 | 0.021 | |

-------------|-----------|-----------|-----------|

Column Total | 57569 | 2374 | 59943 |

| 0.960 | 0.040 | |

-------------|-----------|-----------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test

------------------------------------------------------------

Chi^2 = 60.99267 d.f. = 1 p = 5.728781e-15

Pearson's Chi-squared test with Yates' continuity correction

------------------------------------------------------------

Chi^2 = 60.65958 d.f. = 1 p = 6.784907e-15

> okc %>%

+ with(., table(male, coupled)) %>%

+ chisq.test()

Pearson's Chi-squared test with Yates' continuity correction

data: .

X-squared = 60.66, df = 1, p-value = 6.785e-15

* + 1. Conclusion: Gender is associated with being in a relationship on OkCupid (. On OkCupid, 4.7% of males are in a relationship, whereas that proportion is 3.5% for females.

1. **Odds Ratio**
   1. Odds ratio from a contingency table
      1. There is no good function in base R to compute an odds ratio (that I’m aware of).
      2. We can write a function that returns the odds ratio using the values input from a table (see code).

get.or <- function(table) {

+ or <- table[1]\*table[4]/(table[2]\*table[3])

+ se <- sqrt(1/table[1] + 1/table[2] + 1/table[3] + 1/table[4])

+ upper.95ci <- exp(log(or) + 1.96\*se)

+ lower.95ci <- exp(log(or) - 1.96\*se)

+

+ tibble(or, lower.95ci, upper.95ci)

+ }

> okc %>%

+ with(., table(male, coupled)) %>%

+ get.or()

# A tibble: 1 x 3

or lower.95ci upper.95ci

<dbl> <dbl> <dbl>

1 0.722 0.665 0.783

* + 1. We could also use the epitools package to obtain the odds ratio via the function oddsratio().

> okc %>%

+ with(.,

+ oddsratio(male, coupled))

$data

Outcome

Predictor 0 1 Total

0 22979 1138 24117

1 34590 1236 35826

Total 57569 2374 59943

$measure

odds ratio with 95% C.I.

Predictor estimate lower upper

0 1.0000000 NA NA

1 0.7215298 0.6645838 0.7834219

$p.value

two-sided

Predictor midp.exact fisher.exact chi.square

0 NA NA NA

1 9.325873e-15 9.581349e-15 5.728781e-15

$correction

[1] FALSE

attr(,"method")

[1] "median-unbiased estimate & mid-p exact CI"

* 1. Odds ratio from logistic regression
     1. When there is only one independent variable under consideration, the conclusions obtained from contingency table analysis will be the same as the conclusions reached from a logistic regression.

> glm(coupled ~ male, data = okc, family = binomial) %>% summary()

Call:

glm(formula = coupled ~ male, family = binomial, data = okc)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.3109 -0.3109 -0.2650 -0.2650 2.5949

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.00531 0.03037 -98.961 < 2e-16 \*\*\*

male -0.32638 0.04196 -7.779 7.3e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19983 on 59942 degrees of freedom

Residual deviance: 19923 on 59941 degrees of freedom

(3 observations deleted due to missingness)

AIC: 19927

Number of Fisher Scoring iterations: 6

* + 1. R isn’t as user-friendly with its output and many useful pieces of information need to be computed by hand. This includes things like:
       - The likelihood ratio test vs. the null model
       - The pseudo R-squared
       - The odds ratios

Here, we will manually enter code to compute the odds ratio

> glm(coupled ~ male, data = okc, family = binomial) %>% coef() %>% exp()

(Intercept) male

0.04952348 0.72153395

* + 1. Recall, a 95% confidence interval for the OR of a regression parameter estimate is given as:

1. **Model Assessment**
   1. Model “Error”
      1. Deviance—conceptually, the deviance is the “lack of fit” of a model. A higher deviance indicates a model that doesn’t align well with the data. The deviance is computed as D =   
         -2\*(log likelihood). It is analogous to the SSE (sum of squares error) in OLS regression.
      2. Log likelihood—conceptually, the likelihood indicates how well a model fits. Models with higher likelihood have parameters that align more closely with the data.
      3. In our example, adding BMI to the model reduces the deviance by 60 (from 19983 to 19923).

> glm(coupled ~ male, data = okc, family = binomial) %>% summary()

Call:

glm(formula = coupled ~ male, family = binomial, data = okc)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.3109 -0.3109 -0.2650 -0.2650 2.5949

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.00531 0.03037 -98.961 < 2e-16 \*\*\*

male -0.32638 0.04196 -7.779 7.3e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19983 on 59942 degrees of freedom

Residual deviance: 19923 on 59941 degrees of freedom

(3 observations deleted due to missingness)

AIC: 19927

Number of Fisher Scoring iterations: 6

* 1. The Likelihood Ratio Test
     1. The likelihood ratio test statistic, in terms of the deviance, is:

, where *D0* is the reduced model deviance, *D1* is the full model deviance, and *k* is the difference in the number of parameters between models

* + 1. To compute the likelihood ratio test compared to the null model, we can use the anova() function and specify we want a Chi-Square test, which will compute the p-value for the test statistic given above.

> glm(coupled ~ male, data = okc, family = binomial) %>% anova(test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: coupled

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 59942 19983

male 1 60.036 59941 19923 9.315e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* + 1. Suppose we want to evaluate the addition of “body type” to this model. We could perform the likelihood ratio test comparing Model 1 (male + body\_type) to Model 0 (male). Here, we see that the addition of the dummy variable set for body type statistically significantly improves model fit ().

> anova(couple\_male.m, couple\_male\_body.m, test = "Chisq")

Analysis of Deviance Table

Model 1: coupled ~ male

Model 2: coupled ~ male + body\_type

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 59941 19923

2 59929 19471 12 452.5 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* 1. Pseudo R-Squared
     1. Several measures have been proposed to re-create the measure of R2 for logistic models.
     2. One of the easiest measures, which can be computed by hand, is the McFadden’s R-squared:

Where D0 is the deviance of the null model and D1 is the deviance of the model under consideration. Theoretically, a really good model will have low deviance, which will make D1/D0 small and thus R2 large.

Let’s apply it to the model with just “male” in it. Here, the pseudo R-squared would be 1-(19923/19983) = 0.3%.

* + 1. A list of other methods of computing pseudo R-squared can be found at:  
       <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/>
    2. We can use the DescTools package to get the pseudo R-square values. Here, we compute the McFadden R-square value, and then we compute the Nagelkerke R-squared (another common measure of pseudo-R2).

> PseudoR2(couple\_male.m)

McFadden

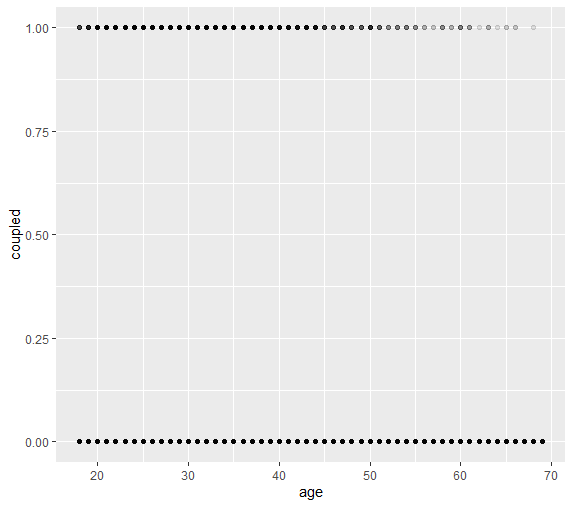
0.003004324

> PseudoR2(couple\_male.m, "Nagelkerke")

Nagelkerke

0.003531102

1. **Prediction**
   1. Suppose we want to determine whether being coupled on OkCupid is related to age. Remember, 1=coupled (married/dating) and 0=single/available. For this example, I removed some outliers in age by filtering age<90.  
      We see that age reduces the likelihood that an individual is coupled. Namely, each one-year increase in age is associated with exp(-0.026) = 0.97 times the odds of being coupled.



> glm(coupled ~ age,

+ data = okc %>%

+ filter(age < 90),

+ family = binomial) %>%

+ summary()

Call:

glm(formula = coupled ~ age, family = binomial, data = okc %>%

filter(age < 90))

Deviance Residuals:

Min 1Q Median 3Q Max

-0.3369 -0.3044 -0.2893 -0.2611 2.8840

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.371419 0.080875 -29.32 <2e-16 \*\*\*

age -0.026055 0.002564 -10.16 <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: 19983 on 59943 degrees of freedom

Residual deviance: 19869 on 59942 degrees of freedom

AIC: 19873

Number of Fisher Scoring iterations: 6

* 1. We can show the best-fit logistic regression equation for π (the probability of being coupled) directly by using ggplot. Remember, the relationship is only linear in the logit, but not linear in terms of probability outcome.

> okc %>%

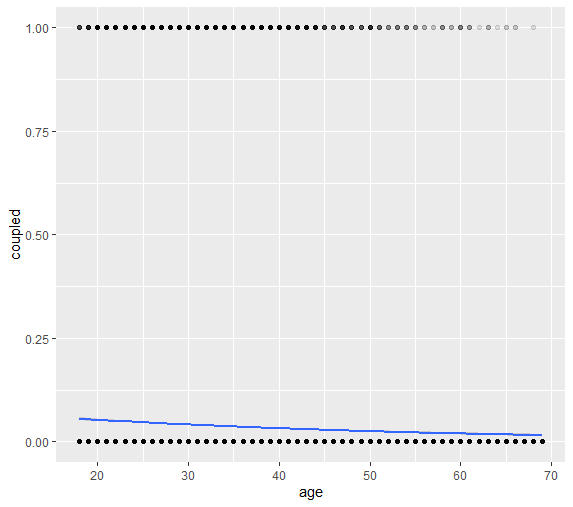
+ filter(age < 90) %>%

+ ggplot(aes(x = age, y = coupled)) +

+ geom\_point(alpha = .1) +

+ geom\_smooth(method = "glm", method.args = list(family = "binomial"))

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



* 1. Recall the predicted value π for any X is given by:
     1. What is the predicted probability of being coupled for somebody 20 years old? 5.3%
     2. What is the predicted probability of being coupled for somebody 40 years old? 3.2%
     3. What is the predicted probability of being coupled for somebody 60 years old? 1.9%
  2. Using the predict() function
     1. We can use the predict() function to obtain the predicted values at these ages:

> predict.glm(coupled\_age.m, tibble(age = c(20, 40, 60)))

1 2 3

-2.892512 -3.413605 -3.934698

* + 1. Uh oh! These numbers are negative—there’s no way they reflect a predicted probability. This is because the predict() function will always first predict the value of the linear predictor. (In this case, it returns the predicted logit.)  
       e.g., -2.892512 = -2.37 – 0.026(20)
    2. To have R to automatically convert these to the predicted probabilities, you must specify type = “response”.

> predict(coupled\_age.m, data.frame(age = c(20, 40, 60)), type = "response")

1 2 3

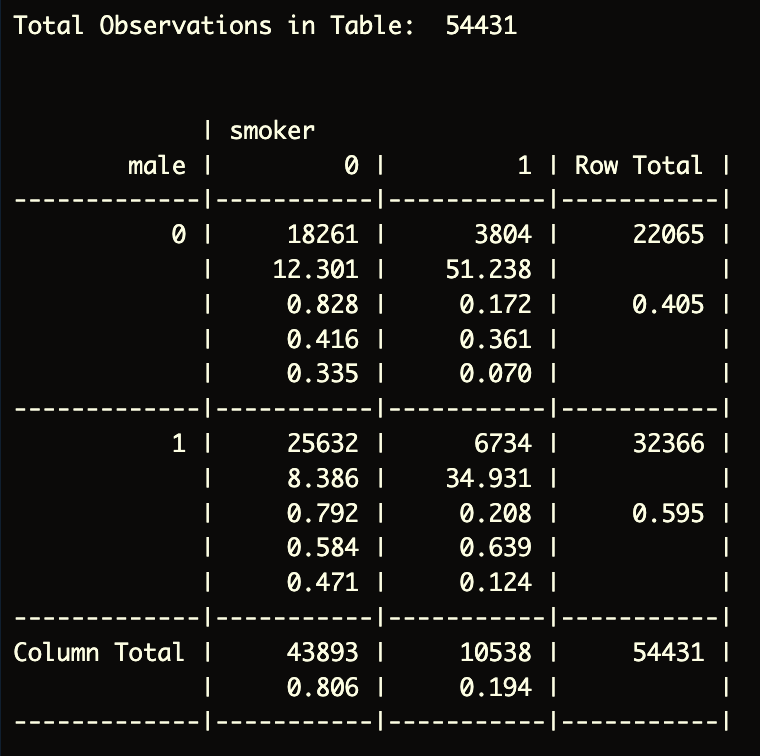
0.05252498 0.03187298 0.01917667

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lab 8 Exercises** |  |  |  |  |  |  |

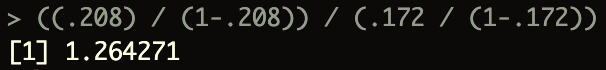
|  |  |
| --- | --- |
| Objective(s): | Use techniques for the analyses of binary variables: contingency tables, odds ratios, logistic regression |
| Datasets Required: | okcprofiles\_cleaned |

Continue to use the OkCupid data set. Suppose a public health agency wants to know about the demographics of individuals who smoke in order to tailor resources to these demographic groups. Examine the effect of gender, age, sexual orientation, and religion on the likelihood of smoking.

1. Univariately, is gender related to smoking?
   1. Produce a contingency table that shows the probability of being a smoker for each level of gender.



* 1. Calculate the odds ratio and confidence interval from this contingency table, and provide a p-value for the relationship.



A screenshot of a computer screen

Description automatically generated

The odds ratio of being a smoker is 1.26 times that for males compared to females. (p<0.001)

* 1. Note: you could also run a univariable logistic regression to retrieve this information, but you do not have to for this exercise.

A computer screen shot of a black screen

Description automatically generated

The odds ratio of being a smoker is exp(0.23204) = 1.26 times that for males compared to females (p<0.01)

1. Univariately, is sexual orientation related to smoking?
   1. Produce a contingency table that shows the probability of being a smoker for those who identified as straight vs. not-straight.

A black and white screen with numbers and lines

Description automatically generated

* 1. Calculate the odds ratio and confidence interval from this contingency table, and provide a p-value for the relationship.

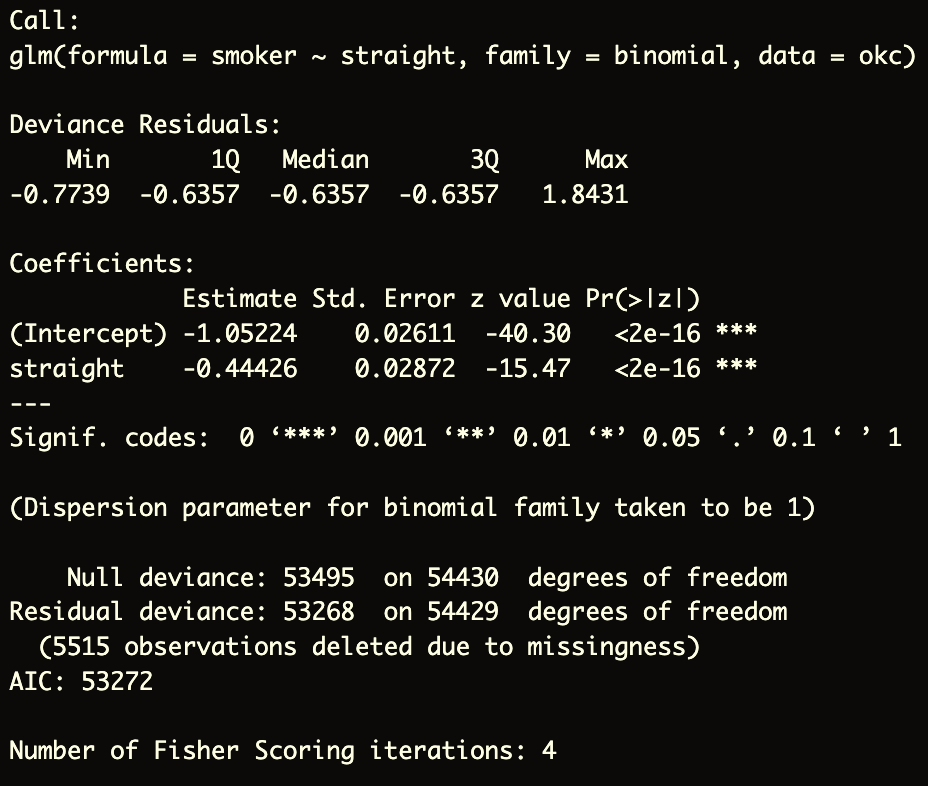
A screenshot of a computer

Description automatically generated

The odds ratio of being a smoker is 0.641 times that for straight people compared to people who are not straight. (p<0.001)

(also can be interpreted as 36 times lower the odds)

* 1. Note: you could also run a univariable logistic regression to retrieve this information, but you do not have to for this exercise.



The odds ratio of being a smoker is exp(-0.44426) = 0.6412987 times that for straight people compared to people who are not straight (p<0.001).

1. Univariately, is age related to smoking?
   1. Produce a visual (e.g., scatter plot, boxplot, etc.) that will convey information about this relationship.

A graph with a blue line

Description automatically generated

A graph with a graph of numbers and a graph with a graph of numbers

Description automatically generated with medium confidence

* 1. Perform a univariable logistic regression with age as the independent variable and smoking as the outcome. Report the parameter estimate for age. What is the interpretation of this parameter estimate? Report the p-value for the relationship between age and smoking.

A screenshot of a computer

Description automatically generated

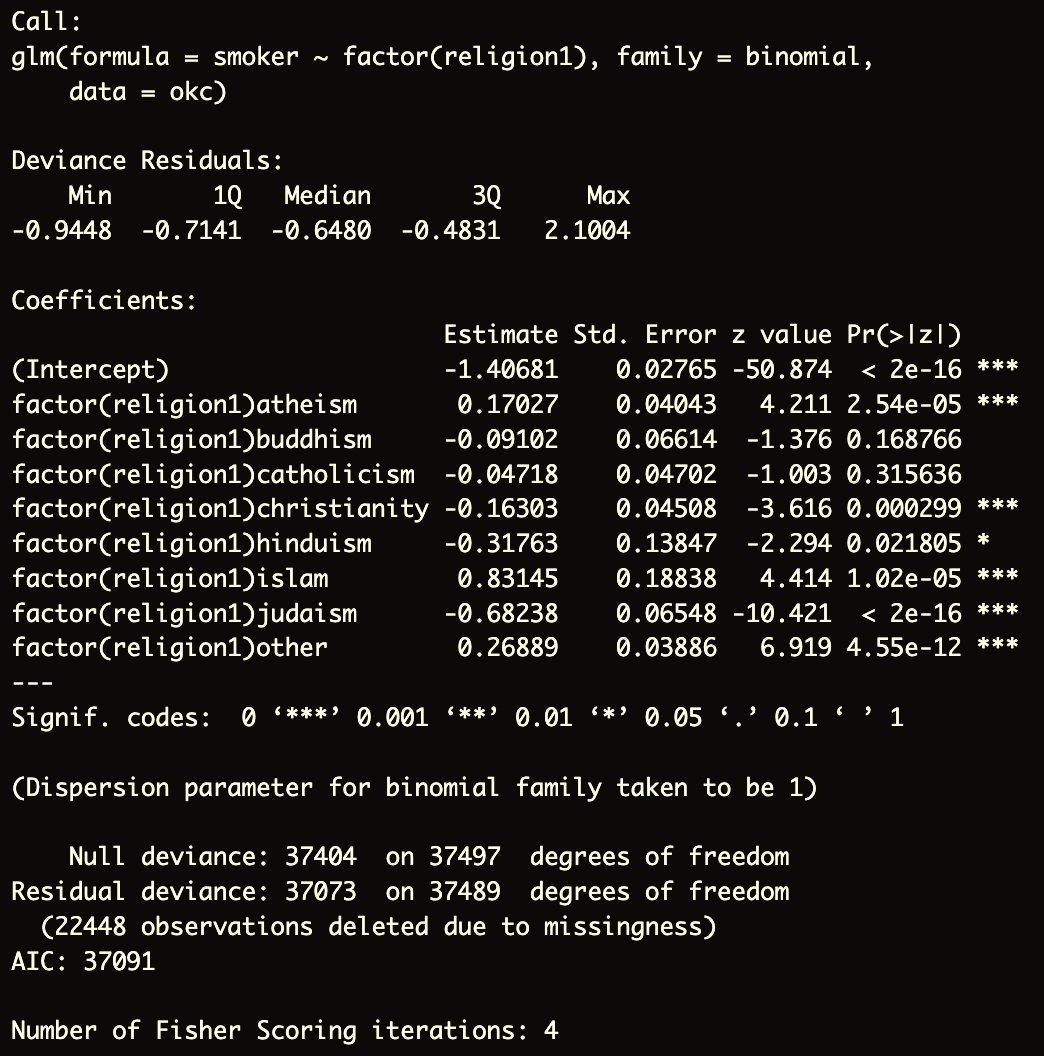
The beta parameter for age is -0.0541 and indicates that a 1-unit increase in age is associated with a exp(-0.0541) = 0.9472881 change in odds for being a smoker (p<0.001).

1. Univariately, is religion related to smoking?
   1. Produce a visual (e.g., scatter plot, boxplot, etc.) that will convey information about this relationship.

A graph with black dots

Description automatically generated

* 1. Perform a logistic regression with religion as a dummy variable set and smoking as the outcome.



* 1. What is the reference group?

Reference group is agnostic

* 1. Select 3 of the coefficients and provide an interpretation for each.

The odds of smoking for someone who is an atheist is exp(0.17027) = 1.185625 times the odds of smoking compared to someone who is agnostic (p<0.001).

The odds of smoking for someone who is a buddhist is exp(-0.091012) = 0.9130068 times the odds of smoking compared to someone who is agnostic (p=0.17).

The odds of smoking for someone who is muslim is exp(0.83145) = 2.296646 times the odds of smoking compared to someone who is agnostic (p<0.001).

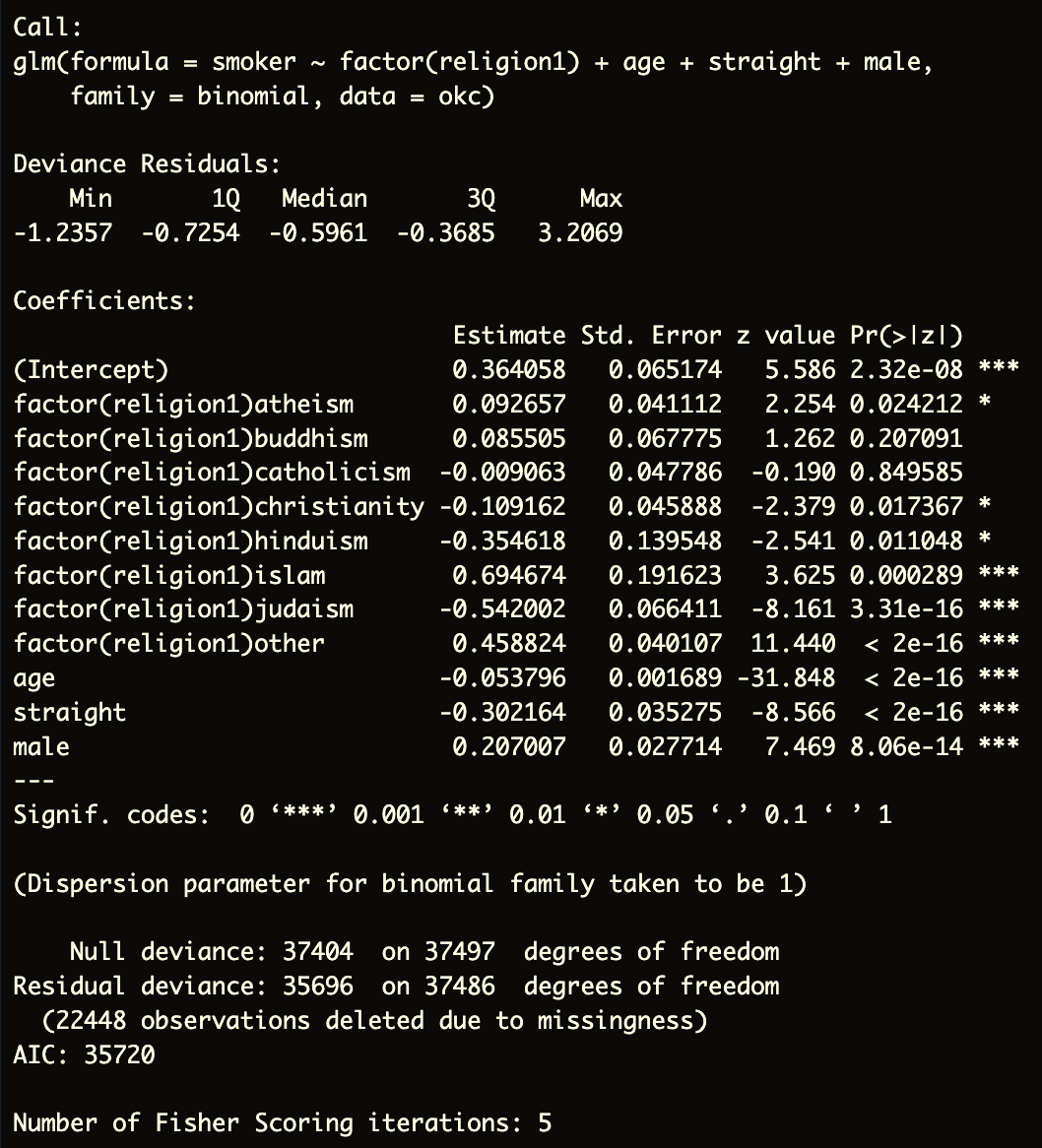
* 1. What is the p-value for the relationship between religion and smoking? (Note: you will need to compute the likelihood ratio test.)

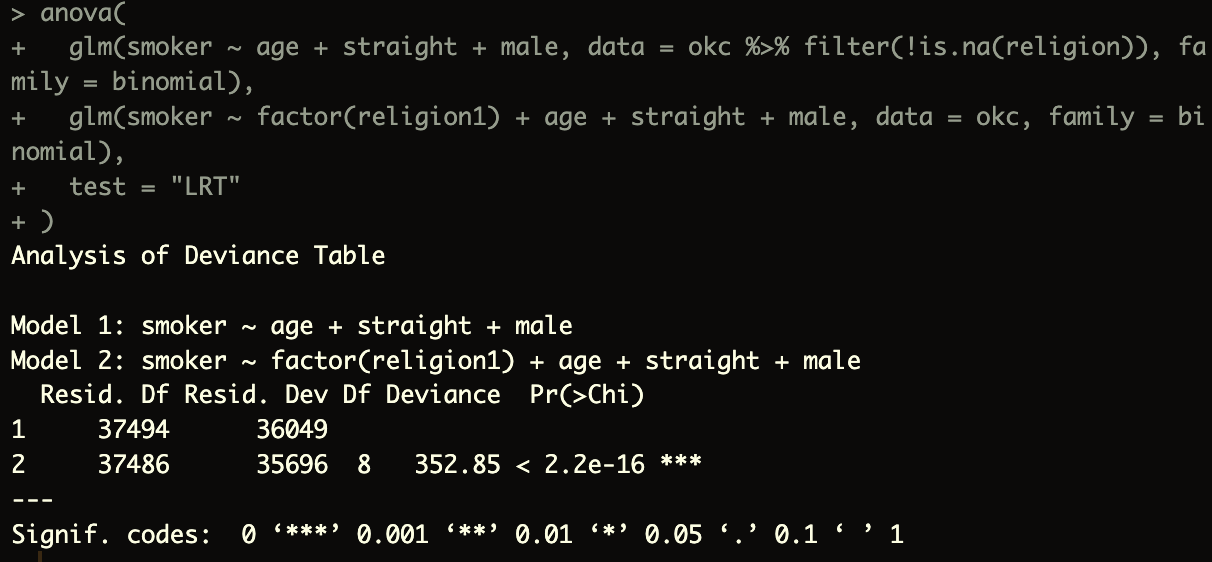
A computer screen with white text

Description automatically generated

p < 0.001

1. Combine all variables into one multivariable model.
   1. Which variables are significant predictors of smoking in this model?





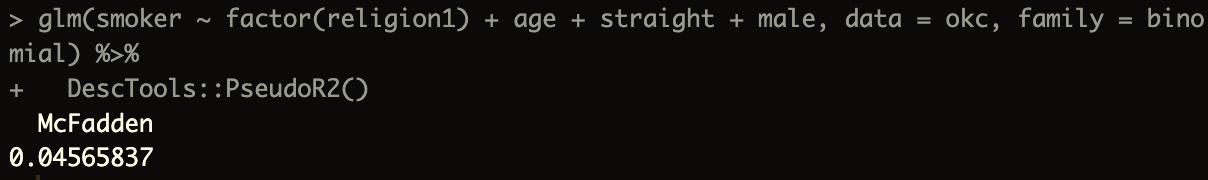
All variables seem to be significant in the model.

* 1. Interpret the beta coefficients for gender and sexual orientation.

The beta coefficient for gender indicates that adjusting for all other variables, the odds for smoking for males is exp(0.207007) = 1.23 increased odds for smoking compared to females.

The beta coefficient for sexual orientation indicates that adjusting for all other variables, the odds for smoking for people who identify as straight is exp(-0.302164) = 0.74 times the odds of smoking for those who do not identify as straight.

* 1. Report the value of the pseudo R2.



* 1. What is the probability of smoking for a straight, 30-year-old, Buddhist female?