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

1.1. Table

1.1.1. 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
```

1.1.2. 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
```

1.2. Xtabs

1.2.1. xtabs(), or cross-tabs, is another way to make a table

1.2.2. You can perform follow-up functions, like prop.table() or chisq.test(), using this method as well.

1.2.3. xtabs can be used to make a table by specifying frequencies of each covariate pattern (see lecture).

1.3. CrossTable

1.3.1. 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 |           |
-------------|-----------|-----------|-----------|
```

1.4. The Chi-Square test

1.4.1. 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
```

1.4.2. Conclusion: Gender is associated with being in a relationship on OkCupid ($\chi^2_1 = 60.99, p < .001$. On OkCupid, 4.7% of males are in a relationship, whereas that proportion is 3.5% for females.

2. Odds Ratio

2.1. Odds ratio from a contingency table

2.1.1. There is no good function in base R to compute an odds ratio (that I'm aware of).

2.1.2. 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
```

2.1.3. 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"
```

2.2. Odds ratio from logistic regression

2.2.1. 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
```

2.2.2. 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
```

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

$$\left(e^{\beta_1 - 1.96SE(\beta_1)}, e^{\beta_1 + 1.96SE(\beta_1)}\right)$$

3. **Model Assessment**
   3.1. Model "Error"

3.1.1. 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.

3.1.2. Log likelihood—conceptually, the likelihood indicates how well a model fits. Models with higher likelihood have parameters that align more closely with the data.

3.1.3. 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
```

3.2. The Likelihood Ratio Test

3.2.1. The likelihood ratio test statistic, in terms of the deviance, is:

$G = D_0 - D_1 \sim \chi^2_k$, where D_0 is the reduced model deviance, D_1 is the full model deviance, and k is the difference in the number of parameters between models

3.2.2. 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
```

3.2.3. 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 ($\chi^2_{12} = 452.5, p < .001$).

```
> 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
```

3.3. Pseudo R-Squared

3.3.1. Several measures have been proposed to re-create the measure of R^2 for logistic models.

3.3.2. One of the easiest measures, which can be computed by hand, is the McFadden's R-squared:

$$R^2_{McFadden} = 1 - \frac{D_1}{D_0}$$

Where D_0 is the deviance of the null model and D_1 is the deviance of the model under consideration. Theoretically, a really good model will have low deviance, which will make D_1/D_0 small and thus R^2 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%.

3.3.3. 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/

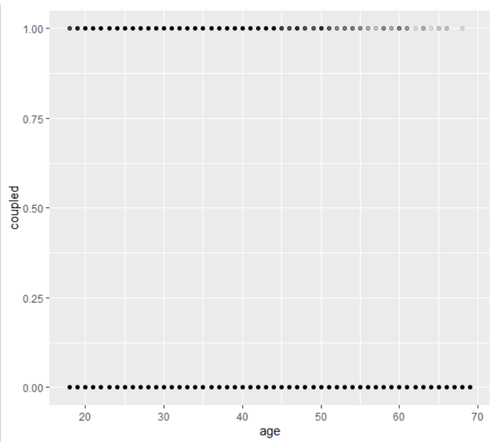
3.3.4. We can use the DescTools package to get the pseudo R-square values. Here, we compute the McFadden R-square value, and then we compute the Nagelkerke R-squared (another common measure of pseudo-R^2).

```
> PseudoR2(couple_male.m)
   McFadden
0.003004324
> PseudoR2(couple_male.m, "Nagelkerke")
 Nagelkerke
0.003531102
```

4. Prediction

4.1. 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.

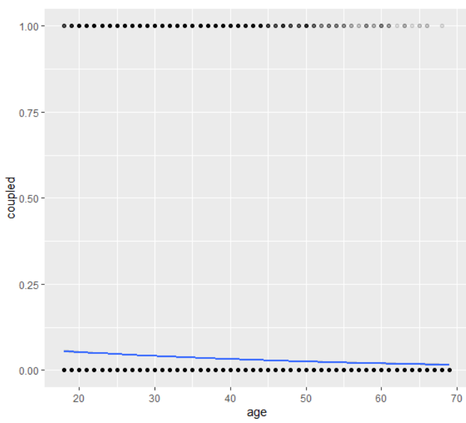



```
> glm(coupled ~ age,
+     data = okc %>%
+       filter(age < 90),
+     family = binomial) %>%
+   summary()

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
```

4.2. 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'
```



4.3. Recall the predicted value π for any X is given by:

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

$$\hat{\pi} = \frac{e^{-2.37-0.026X_1}}{1+e^{-2.37-0.026X_1}}$$

4.3.1. What is the predicted probability of being coupled for somebody 20 years old? 5.3%

$$\frac{e^{-2.37-0.026(20)}}{1+e^{-2.37-0.026(20)}}$$

4.3.2. What is the predicted probability of being coupled for somebody 40 years old? 3.2%

4.3.3. What is the predicted probability of being coupled for somebody 60 years old? 1.9%

4.4. Using the predict() function

4.4.1. 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 
```

4.4.2. 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)

4.4.3. 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?

   a. Produce a contingency table that shows the probability of being a smoker for each level of gender.

```
Total Observations in Table:  54431


             | smoker
        male |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |     18261 |      3804 |     22065 |
             |    12.301 |    51.238 |           |
             |     0.828 |     0.172 |     0.405 |
             |     0.416 |     0.361 |           |
             |     0.335 |     0.070 |           |
-------------|-----------|-----------|-----------|
           1 |     25632 |      6734 |     32366 |
             |     8.386 |    34.931 |           |
             |     0.792 |     0.208 |     0.595 |
             |     0.584 |     0.639 |           |
             |     0.471 |     0.124 |           |
-------------|-----------|-----------|-----------|
Column Total |     43893 |     10538 |     54431 |
             |     0.806 |     0.194 |           |
-------------|-----------|-----------|-----------|
```

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

```
> ((.208) / (1-.208)) / (.172 / (1-.172))
[1] 1.264271
```

```
$data
        Outcome
Predictor     0     1 Total
    0     18261  3804 22065
    1     25632  6734 32366
    Total 43893 10538 54431

$measure
        odds ratio with 95% C.I.
Predictor estimate    lower    upper
       0 1.000000       NA       NA
       1 1.261149 1.206867 1.318046

$p.value
        two-sided
Predictor midp.exact fisher.exact  chi.square
       0         NA           NA          NA
       1          0 3.065375e-25 4.78509e-25
```

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

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

```
Call:
glm(formula = smoker ~ male, family = binomial, data = okc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6830  -0.6830  -0.6152  -0.6152   1.8751

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.56871    0.01782  -88.02   <2e-16 ***
male         0.23204    0.02248   10.32   <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: 53495  on 54430  degrees of freedom
Residual deviance: 53387  on 54429  degrees of freedom
  (5515 observations deleted due to missingness)
AIC: 53391

Number of Fisher Scoring iterations: 4
```

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

2. Univariately, is sexual orientation related to smoking?

   a. Produce a contingency table that shows the probability of being a smoker for those who identified as straight vs. not-straight.

```
             | straight
        male |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      3584 |     20533 |     24117 |
             |    15.567 |     2.516 |           |
             |     0.149 |     0.851 |     0.402 |
-------------|-----------|-----------|-----------|
           1 |      4756 |     31070 |     35826 |
             |    10.479 |     1.694 |           |
             |     0.133 |     0.867 |     0.598 |
-------------|-----------|-----------|-----------|
Column Total |      8340 |     51603 |     59943 |
-------------|-----------|-----------|-----------|
```

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

```
$data
         Outcome
Predictor     0     1 Total
    0      5668  1979  7647
    1     38225  8559 46784
    Total 43893 10538 54431

$measure
         odds ratio with 95% C.I.
Predictor  estimate     lower     upper
        0 1.0000000        NA        NA
        1 0.6412761 0.6062781 0.6785226

$p.value
         two-sided
Predictor midp.exact fisher.exact   chi.square
        0         NA           NA           NA
        1          0  2.15655e-51 1.308502e-54

$correction
[1] FALSE

attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
```

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)

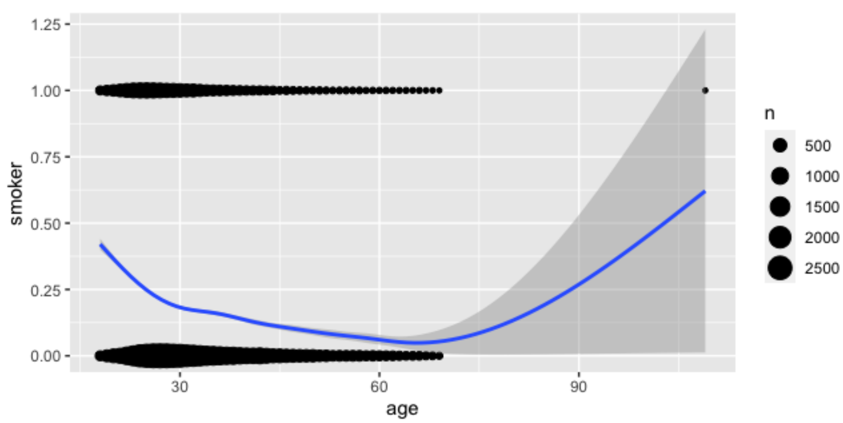
c. Note: you could also run a univariable logistic regression to retrieve this information, but you do not have to for this exercise.

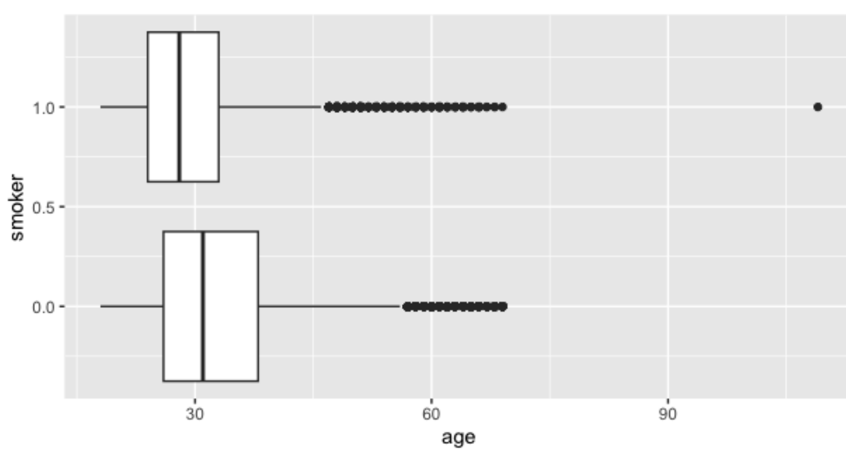
```
Call:
glm(formula = smoker ~ straight, family = binomial, data = okc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7739  -0.6357  -0.6357  -0.6357   1.8431

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.05224    0.02611  -40.30   <2e-16 ***
straight    -0.44426    0.02872  -15.47   <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: 53495  on 54430  degrees of freedom
Residual deviance: 53268  on 54429  degrees of freedom
  (5515 observations deleted due to missingness)
AIC: 53272

Number of Fisher Scoring iterations: 4
```

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 ($p<0.001$).

3. Univariately, is age related to smoking?

   a. Produce a visual (e.g., scatter plot, boxplot, etc.) that will convey information about this relationship.





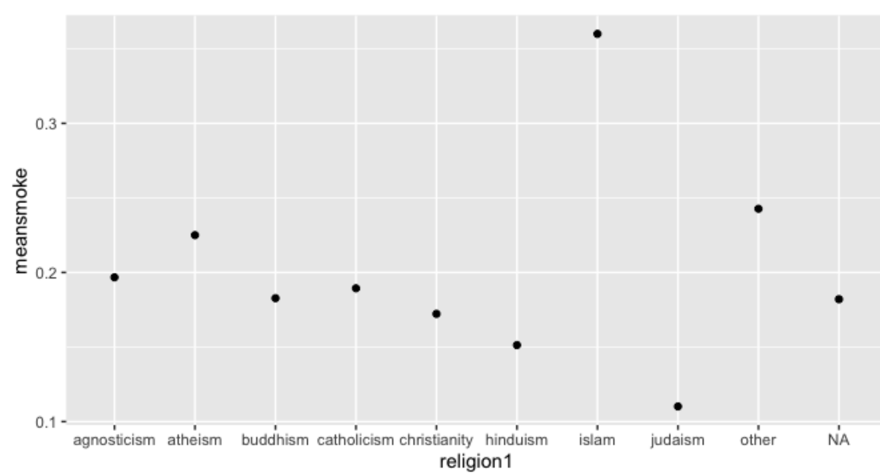
b. Perform a univariable logistic regression with age as the independent variable and smoking as the outcome. Report the β parameter estimate for age. What is the interpretation of this parameter estimate? Report the p-value for the relationship between age and smoking.

```
Call:
glm(formula = smoker ~ age, family = binomial, data = okc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8923  -0.7248  -0.6115  -0.3967   3.3606

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.259315   0.044871   5.779 7.51e-09 ***
age         -0.054152   0.001456 -37.198  < 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: 53495  on 54430  degrees of freedom
Residual deviance: 51830  on 54429  degrees of freedom
  (5515 observations deleted due to missingness)
AIC: 51834

Number of Fisher Scoring iterations: 4
```

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).

4. Univariately, is religion related to smoking?

a. Produce a visual (e.g., scatter plot, boxplot, etc.) that will convey information about this relationship.



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

```
Call:
glm(formula = smoker ~ factor(religion1), family = binomial,
    data = okc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9448  -0.7141  -0.6480  -0.4831   2.1004

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.40681    0.02765 -50.874  < 2e-16 ***
factor(religion1)atheism       0.17027    0.04043   4.211 2.54e-05 ***
factor(religion1)buddhism     -0.09102    0.06614  -1.376 0.168766
factor(religion1)catholicism  -0.04718    0.04702  -1.003 0.315636
factor(religion1)christianity -0.16303    0.04508  -3.616 0.000299 ***
factor(religion1)hinduism     -0.31763    0.13847  -2.294 0.021805 *
factor(religion1)islam         0.83145    0.18838   4.414 1.02e-05 ***
factor(religion1)judaism      -0.68238    0.06548 -10.421  < 2e-16 ***
factor(religion1)other         0.26889    0.03886   6.919 4.55e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37404  on 37497  degrees of freedom
Residual deviance: 37073  on 37489  degrees of freedom
  (22448 observations deleted due to missingness)
AIC: 37091

Number of Fisher Scoring iterations: 4
```

c. What is the reference group?

Reference group is agnostic

d. Select 3 of the β 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 ($p<0.001$).

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

```
> glm(smoker ~ factor(religion1), data = okc, family = binomial) %>% anova(test =
"LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: smoker

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              37497      37404
factor(religion1)  8   330.81     37489      37073 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$p < 0.001$

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

```
Call:
glm(formula = smoker ~ factor(religion1) + age + straight + male,
    family = binomial, data = okc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2357  -0.7254  -0.5961  -0.3685   3.2069

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   0.364058   0.065174   5.586 2.32e-08 ***
factor(religion1)atheism      0.092657   0.041112   2.254 0.024212 *
factor(religion1)buddhism     0.085505   0.067775   1.262 0.207091
factor(religion1)catholicism -0.009063   0.047786  -0.190 0.849585
factor(religion1)christianity -0.109162  0.045888  -2.379 0.017367 *
factor(religion1)hinduism    -0.354618   0.139548  -2.541 0.011048 *
factor(religion1)islam        0.694674   0.191623   3.625 0.000289 ***
factor(religion1)judaism     -0.542002   0.066411  -8.161 3.31e-16 ***
factor(religion1)other        0.458824   0.040107  11.440  < 2e-16 ***
age                          -0.053796   0.001689 -31.848  < 2e-16 ***
straight                     -0.302164   0.035275  -8.566  < 2e-16 ***
male                          0.207007   0.027714   7.469 8.06e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37404  on 37497  degrees of freedom
Residual deviance: 35696  on 37486  degrees of freedom
  (22448 observations deleted due to missingness)
AIC: 35720

Number of Fisher Scoring iterations: 5
```

```
> anova(
+   glm(smoker ~ age + straight + male, data = okc %>% filter(!is.na(religion)), fa
mily = binomial),
+   glm(smoker ~ factor(religion1) + age + straight + male, data = okc, family = bi
nomial),
+   test = "LRT"
+ )
Analysis of Deviance Table

Model 1: smoker ~ age + straight + male
Model 2: smoker ~ factor(religion1) + age + straight + male
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     37494      36049
2     37486      35696  8   352.85 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All variables seem to be significant in the model.

b. 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.

c. Report the value of the pseudo R^2.

```
> glm(smoker ~ factor(religion1) + age + straight + male, data = okc, family = bino
mial) %>%
+   DescTools::PseudoR2()
  McFadden
0.04565837
```

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

$$\hat{Y} = 0.364058 + 0.085505X_{buddhist} - 0.302164X_{straight} - 0.053796X_{age} + 0.207007X_{male} + 0(\dots)$$

$$\hat{Y} = 0.364058 + 0.085505(1) - 0.302164(1) - 0.053796(30) + 0.207007(0)$$

$$\hat{Y} = -1.466481$$

$$\hat{\pi} = \frac{e^{-1.466481}}{1 + e^{-1.466481}} = 0.1874781$$