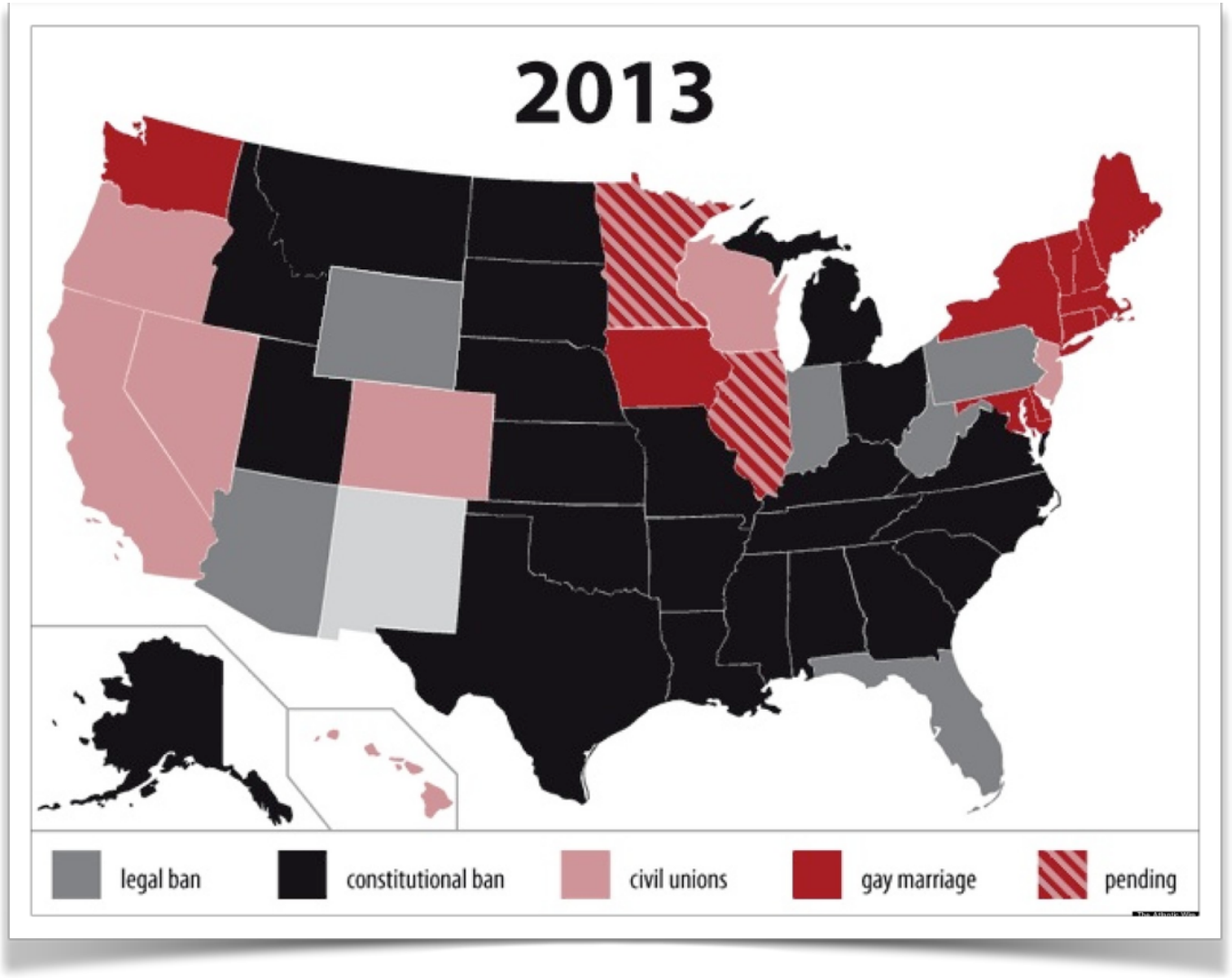
Logistic Regression

Andrew Zieffler
Department of Educational Psychology

Research Question

Which demographic factors play a role in shaping public opinion regarding same-sex marriage?



Reading and Examining the Data

```
# Read in the data
gay = read.csv(file = "http://www.tc.umn.edu/~zief0002/Data/Gay-Marriage.csv")


# Examine data
head(gay)
```

```
  id marriage attend denom gender ideology orientation friends educ region race age
1  1        3      2     1      2        0           1       5   12      3    1  58
2  2        3      1     1      2        5           1       5   16      3    1  39
3  3        2      4     2      1        3           1       1   12      3    1  50
4  4        3      2     1      1        6           1       5   16      3    1  72
5  5        3      5     7      2        5           1       5   14      3    1  71
6  6        3      1     1      2        4           1       1   16      3    1  66
```

marriage (Response)

attend (Focal predictor)

- **marriage:** Respondent supports gay marriage
    1. *Yes*
    2. *No, but civil unions yes*
    3. *No*
- **attend:** How often does the respondent attend religious services?
    1. *Every week;*
    2. *Almost every week;*
    3. *Once or twice a month*
    4. *Few times a year*
    5. *Never*

```r
# Look at structure of the data and variables
str(gay)
```

```
'data.frame':   1746 obs. of  12 variables:
 $ id          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ marriage    : int  3 3 2 3 3 3 3 3 3 2 ...
 $ attend      : int  2 1 4 2 5 1 1 1 1 3 ...
 $ denom       : int  1 1 2 1 7 1 1 1 1 1 ...
 $ gender      : int  2 2 1 1 2 2 1 2 1 2 ...
 $ ideology    : int  0 5 3 6 5 4 4 6 0 3 ...
 $ orientation: int  1 1 1 1 1 1 1 1 1 1 ...
 $ friends     : int  5 5 1 5 5 1 5 5 5 1 ...
 $ educ        : int  12 16 12 16 14 16 14 14 16 15 ...
 $ region      : int  3 3 3 3 3 3 3 3 3 3 ...
 $ race        : int  1 1 1 1 1 1 1 1 2 1 ...
 $ age         : int  58 39 50 72 71 66 56 40 55 84 ...
```

Several analytical questions

- Should **marriage** be turned into a factor?
- Should categories be merged?
- Can we treat **attend** as quantitative? Or should it be treated as categorical?
- If **attend** is treated as quantitative, should the values be reverse-coded?

All variables are being treated as quantitative variables (integers)

- According to the codebook **marriage is categorical** with three levels
    1. *Yes*
    2. *No, but civil unions yes*
    3. *No*
- According to the codebook **attend is ordinal** with five levels
    1. *Every week;*
    2. *Almost every week;*
    3. *Once or twice a month*
    4. *Few times a year*
    5. *Never*

```
# Examine marriage
summary(gay$marriage)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.058   3.000   3.000
```

Values look within the bounds specified in codebook

```
> library(gmodels)
> CrossTable(gay$marriage, format = "SPSS")
```

```
Total Observations in Table:  1746

          |         1  |         2  |         3  |
          |------------|------------|------------|
          |       594  |       456  |       696  |
          |    34.021% |    26.117% |    39.863% |
          |------------|------------|------------|
```

Based on the 1,746 responses to the item...

- 594 (34%) support gay marriage
- 456 (26%) do not support gay marriage, but do support civil unions
- 696 (40%) do not support gay marriage nor civil unions

Based on the research question...

- Should **marriage** be turned into a factor?
- Should categories be merged?

Based on the research question and how "support" has been previously defined in the literature...

- We will **merge** categories 2 and 3 into a single category of "non-support"
- With only two categories (support and non-support) we will use **dummy coding**

```
> gay$support = ifelse(gay$marriage == 1, 1, 0)
```

New variable, **support**

- 1 if the respondent supports gay marriage (marriage = 1)
- 0 if the respondent does nor support gay marriage (marriage = 2 or marriage = 3)

```
> CrossTable(gay$support, format = "SPSS")
```

```
Total Observations in Table:  1746

          |        0  |        1  |
          |-----------|-----------|
          |     1152  |      594  |
          |   65.979% |   34.021% |
          |-----------|-----------|
```

34% of the surveyed respondents support gay marriage

```
> summary(gay$support)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3402  1.0000  1.0000
```

```
# Examine focal predictor
> summary(gay$attend)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   3.000   2.861   4.000   5.000
```

Values look within the bounds specified in codebook

```
> CrossTable(gay$attend, format = "SPSS")
```

```
Total Observations in Table:  1746

          |         1  |         2  |         3  |         4  |         5  |
          |------------|------------|------------|------------|------------|
          |       514  |       227  |       351  |       296  |       358  |
          |    29.439% |    13.001% |    20.103% |    16.953% |    20.504% |
          |------------|------------|------------|------------|------------|
```

Based on the 1,746 responses to the item...

- 514 (29%) attend religious services every week
- 227 (13%) attend religious services almost every week
- 351 (20%) attend religious services once or twice a month
- 296 (17%) attend religious services a few times a year
- 358 (21%) never attend religious services

How does the support (or non-support) of gay marriage play out within these five categories?

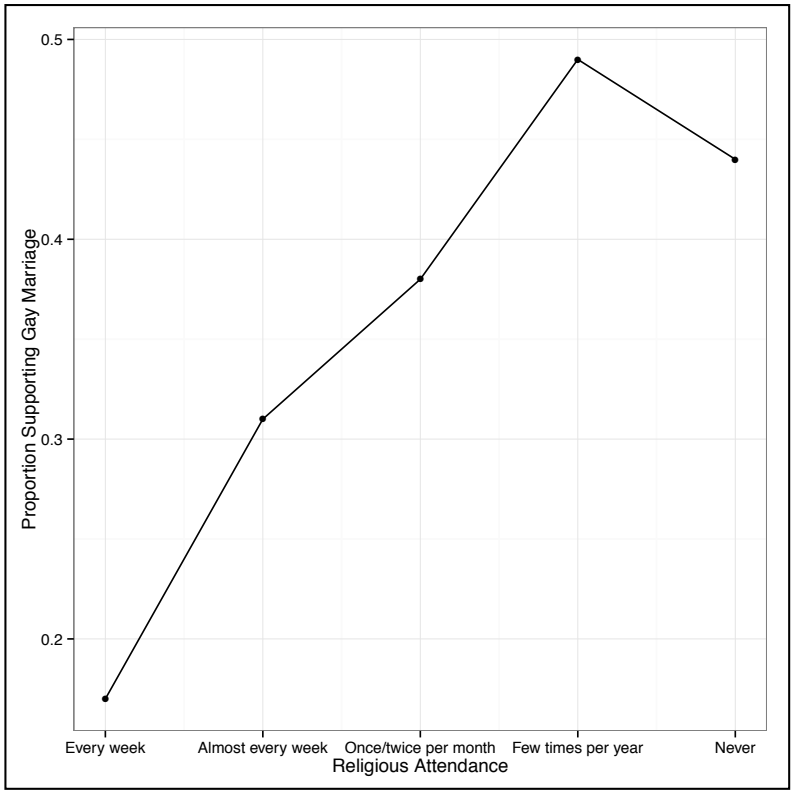
```
# Examine response conditioned on predictor
> CrossTable(gay$support, gay$attend, format = "SPSS")


Total Observations in Table:  1746

             | gay$attend
 gay$support |         1  |         2  |         3  |         4  |         5  | Row Total  |
-------------|------------|------------|------------|------------|------------|------------|
           0 |       426  |       157  |       219  |       151  |       199  |      1152  |
             |    22.250  |     0.349  |     0.684  |    10.048  |     5.861  |            |
             |    36.979% |    13.628% |    19.010% |    13.108% |    17.274% |    65.979% |
             |    82.879% |    69.163% |    62.393% |    51.014% |    55.587% |            |
             |    24.399% |     8.992% |    12.543% |     8.648% |    11.397% |            |
-------------|------------|------------|------------|------------|------------|------------|
           1 |        88  |        70  |       132  |       145  |       159  |       594  |
             |    43.151  |     0.676  |     1.327  |    19.487  |    11.366  |            |
             |    14.815% |    11.785% |    22.222% |    24.411% |    26.768% |    34.021% |
             |    17.121% |    30.837% |    37.607% |    48.986% |    44.413% |            |
             |     5.040% |     4.009% |     7.560% |     8.305% |     9.107% |            |
-------------|------------|------------|------------|------------|------------|------------|
Column Total |       514  |       227  |       351  |       296  |       358  |      1746  |
             |    29.439% |    13.001% |    20.103% |    16.953% |    20.504% |            |
-------------|------------|------------|------------|------------|------------|------------|
```

- **Of the 514 respondents who attend religious services every week**
  - 88 (17%) support gay marriage
  - 426 (83%) do not support gay marriage
- **Of the 227 respondents who attend religious services almost every week**
  - 70 (31%) support gay marriage
  - 157 (69%) do not support gay marriage
- **Of the 351 respondents who attend religious services once or twice a month**
  - 132 (38%) support gay marriage
  - 219 (62%) do not support gay marriage
- **Of the 296 respondents who attend religious services a few times a year**
  - 145 (49%) support gay marriage
  - 151 (51%) do not support gay marriage
- **Of the 358 respondents who never attend religious services**
  - 159 (44%) support gay marriage
  - 199 (56%) do not support gay marriage

- **Support of gay marriage seems related to frequency of religious attendance. *(There seems to be more support for gay marriage amongst respondents who attend religious services less frequently*.)**
  - 17% support gay marriage (every week)
  - 31% support gay marriage (almost every week)
  - 38% support gay marriage (once/twice a month)
  - 49% support gay marriage (few times a year)
  - 44% support gay marriage (never)



Analytically...

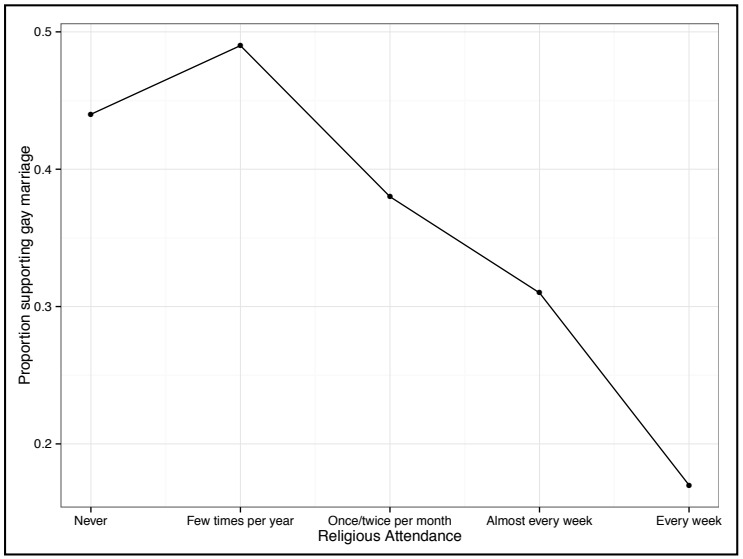
- Can we treat **attend** as quantitative? Or should it be treated as categorical?
- If **attend** is treated as quantitative, should the values be reverse-coded?

Reverse coding = 5 – Original coding

```
# Reverse code attendance
> gay$attend2 = 5 - gay$attend
```

Attendance	Original Coding	Reverse Coding
Every week	1	4
Almost every week	2	3
Once/twice per month	3	2
Few times per year	4	1
Never	5	0

Good practice to make one category 0



	id	marriage	attend	denom	gender	ideology	orientation	friends	educ	region	race	age	support	attend2
1	1	3	2	1	2	0	1	5	12	3	1	58	0	3
2	2	3	1	1	2	5	1	5	16	3	1	39	0	4
3	3	2	4	2	1	3	1	1	12	3	1	50	0	1
4	4	3	2	1	1	6	1	5	16	3	1	72	0	3
5	5	3	5	7	2	5	1	5	14	3	1	71	0	0
6	6	3	1	1	2	4	1	1	16	3	1	66	0	4

```
> glm.a <- glm(support ~ attend2, data = gay, family = binomial(link = "logit"))
> summary(glm.a)

Call:
glm(formula = support ~ attend2, family = binomial(link = "logit"),
    data = gay)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1926  -0.9140  -0.6791   1.1623   1.7777

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03578    0.08463   0.423    0.672
attend2     -0.34635    0.03520  -9.838   <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: 2239.0  on 1745  degrees of freedom
Residual deviance: 2136.9  on 1744  degrees of freedom
AIC: 2140.9

Number of Fisher Scoring iterations: 4
```

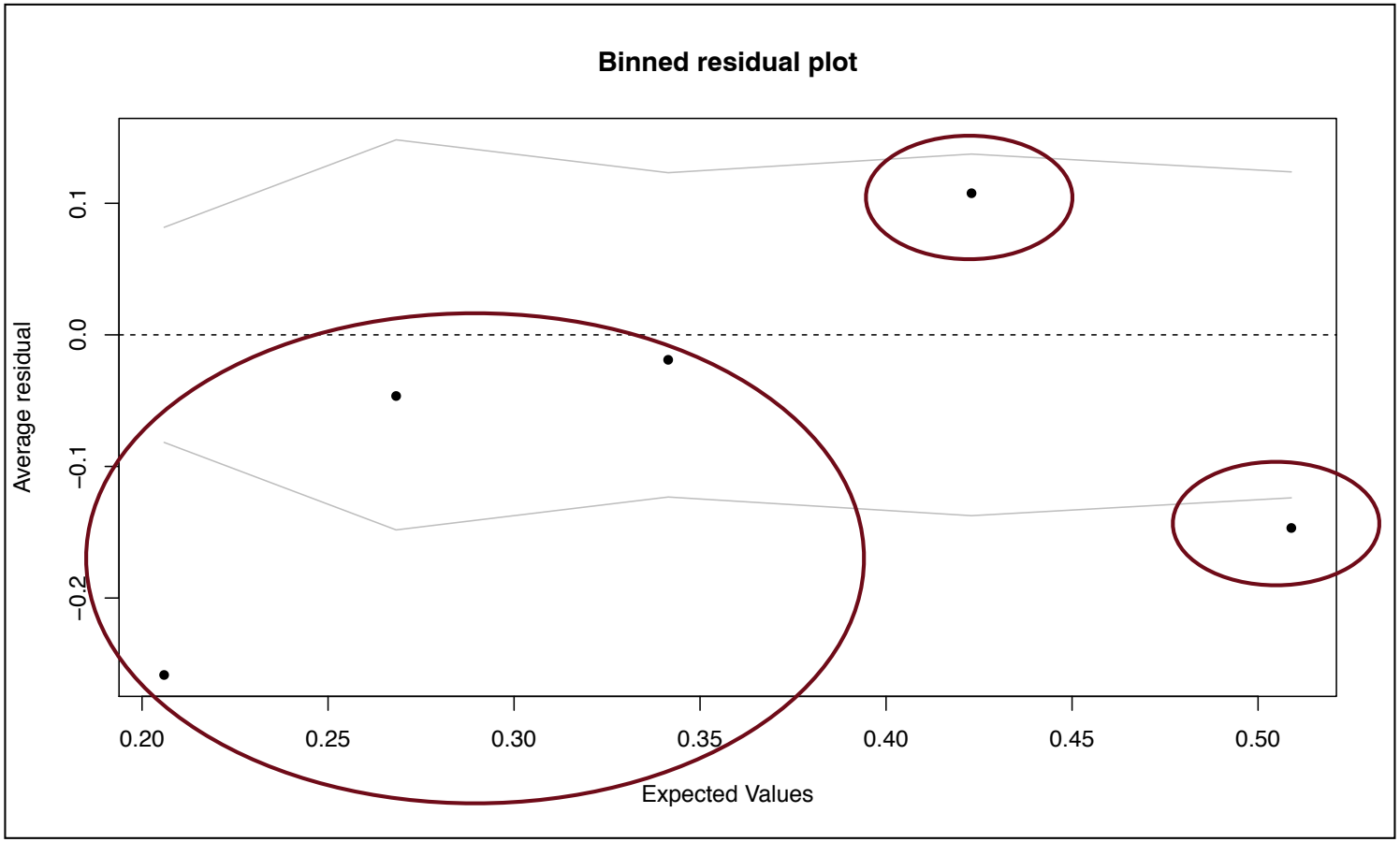
Parameter estimates from the logistic regression (logits/log-odds)

- The intercept, 0.04, is the predicted log-odds of supporting gay marriage for Americans who never attend religious services
- The slope, –0.35, is the predicted change in the log-odds of supporting gay marriage for a one-unit change in support.

Parameter estimates from the logistic regression (odds)

- The intercept, $e^{0.04} = 1.04$, is the **predicted odds** of supporting gay marriage for Americans who *never attend religious services*.
  - ✓ The reciprocal, 1/1.04 = 0.96, gives the **odds of not supporting gay marriage** for Americans who never attend religious services.
- A one-unit change in attendance of religious services is associated with a decrease in the odds of supporting gay marriage by a factor 0.70 ($e^{-0.35}$).

```
# Binned residual plot
> library(arm)
> binnedplot(x = fitted(glm.a), y = resid(glm.a))
```



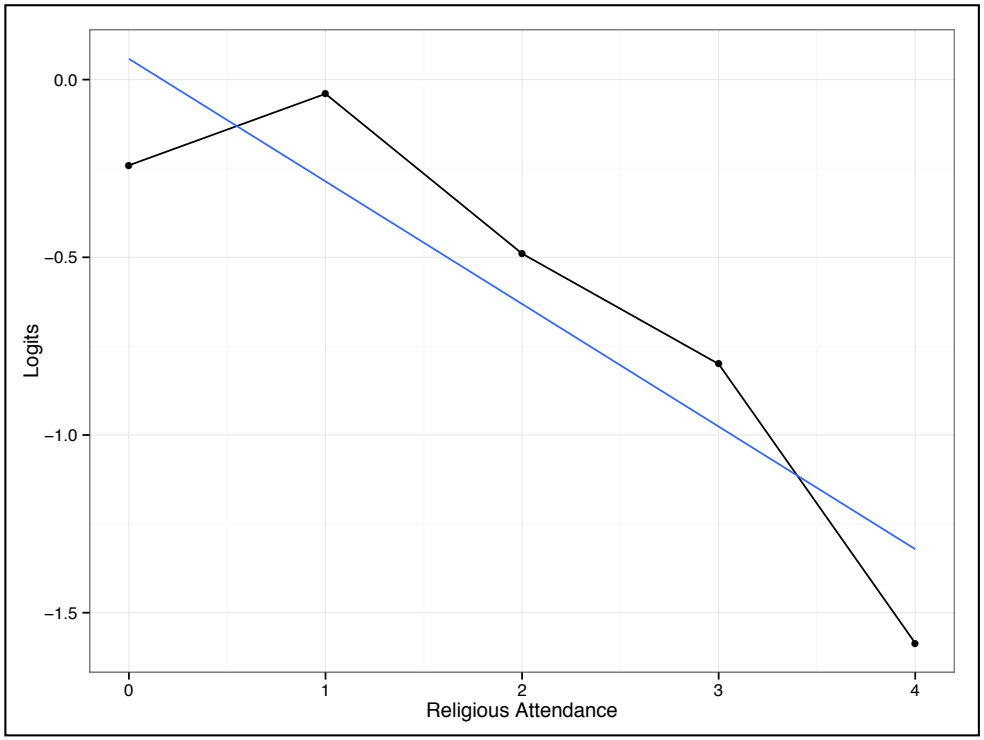
The binned residual plot shows systematic under- and over-prediction across the fitted values. This might suggest model mis-fit...

```
# Create data frame with predictor and logits
> new = data.frame(
      attend = 5 - c(1, 2, 3, 4, 5),
      prop.support = c(0.17, 0.31, 0.38, 0.49, 0.44),
      logits = log(c(0.17, 0.31, 0.38, 0.49, 0.44) / (1 - c(0.17, 0.31, 0.38, 0.49, 0.44)))
      )


# Examine quadratic relationship
> ggplot(data = new, aes(x = attend, y = logits)) +
      geom_point() +
      geom_line() +
      geom_smooth(method = "lm", se = FALSE) +
      theme_bw()
```

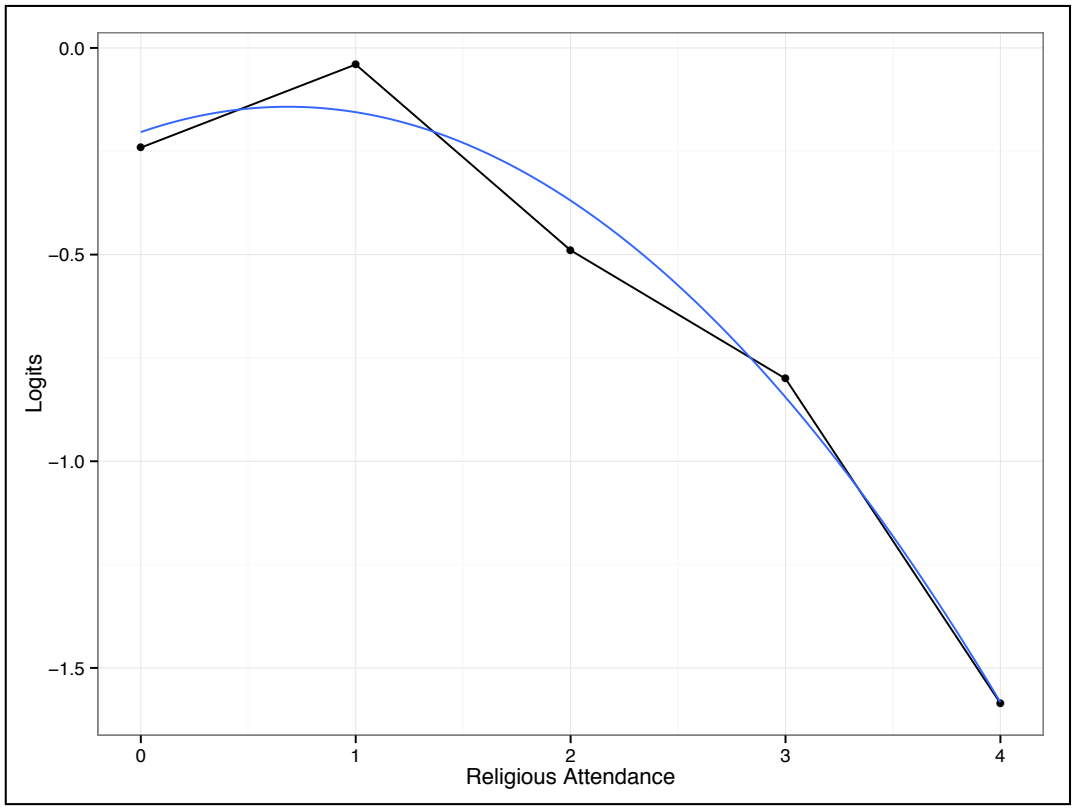
Logit model assumes a linear relationship between the predictor and the log-odds

Does that appear to be the case?



We may want to examine a quadratic relationship between religious attendance and the logits

```
# Examine quadratic relationship
> ggplot(data = new, aes(x = attend, y = logits)) +
    geom_point() +
    geom_line() +
    geom_smooth(method = "lm", se = FALSE, formula = y ~ poly(x, 2)) +
    theme_bw()
```



This seems to fit much better!

Recall that quadratic models are interaction models. The quadratic term is an interaction between the predictor and itself.

Here we are positing that the relationship between religious attendance and the log-odds of supporting gay marriage depends on the level of religious attendance!

```
> glm.b <- glm(support ~ attend2 + I(attend2 ^ 2), data = gay, family = binomial(link = "logit"))
> summary(glm.b)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.18792    0.10124  -1.856   0.0634 .
attend2       0.14007    0.12473   1.123   0.2614
I(attend2^2) -0.12198    0.03015  -4.046 5.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2239.0  on 1745  degrees of freedom
Residual deviance: 2120.5  on 1743  degrees of freedom
AIC: 2126.5

Number of Fisher Scoring iterations: 4
```

The quadratic term is statistically significant and should be included.

- Even though the linear term is not statistically significant, it needs to remain in the model since it is a component of the significant interaction.
  - ✓ As such, it would not be interpreted
- The quadratic term implies the relationship between religious attendance and the log-odds of supporting gay marriage is complex (the rate of change is not constant across values of the predictor).
  - ✓ The term is negative, which implies an upside-down "U"-shaped relationship
  - ✓ To best interpret the relationship between religious attendance and the log-odds of supporting gay marriage, we should create a plot

Parameter estimates from the logistic regression (probabilities)

- The **predicted probability** of supporting gay marriage for Americans who never attend religious services is 0.51.

- The **slope is not directly interpretabl**e in terms of the predicted change in probability
  - ✓ Recall the change in probability is a function of the probability (non-linear change), namely $\beta_1\pi_i(1-\pi_i)$
  - ✓ Thus, the predicted change in probability (effect of religious service attendance) is dependent on where the applicant is on the religious service attendance spectrum.

Table 1
Frequency (Percentage) Supporting Gay Marriage Conditioned on Frequency of Attendance at Religious Services (N=1,746)

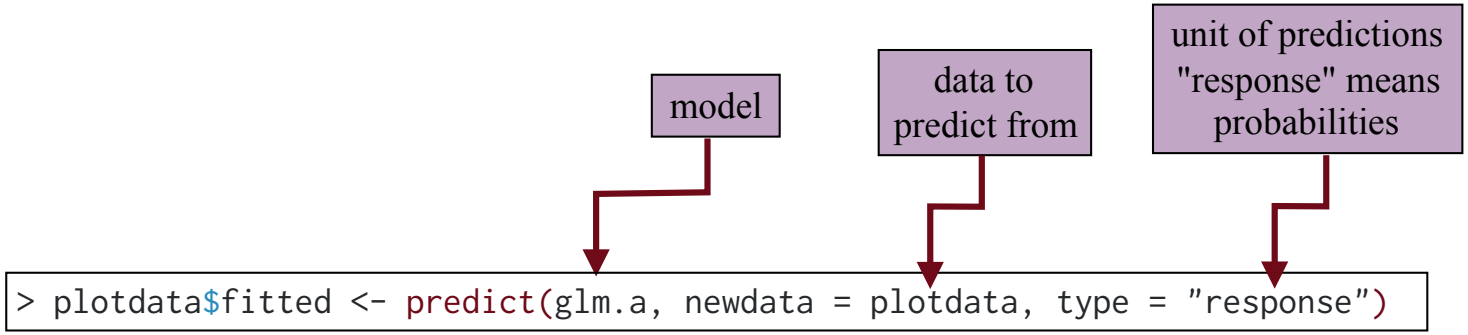
	Supports gay marriage	
Attends religious services	Yes	No
Never	159 (44%)	199 (56%)
Few times per year	145 (49%)	151 (51%)
Once/twice per month	132 (38%)	219 (62%)
Almost every week	70 (31%)	157 (69%)
Every week	88 (17%)	426 (83%)
	594 (34%)	1152 (66%)

Plotting the Fitted Model

Step 1: Create a data frame of the predictors in the model.

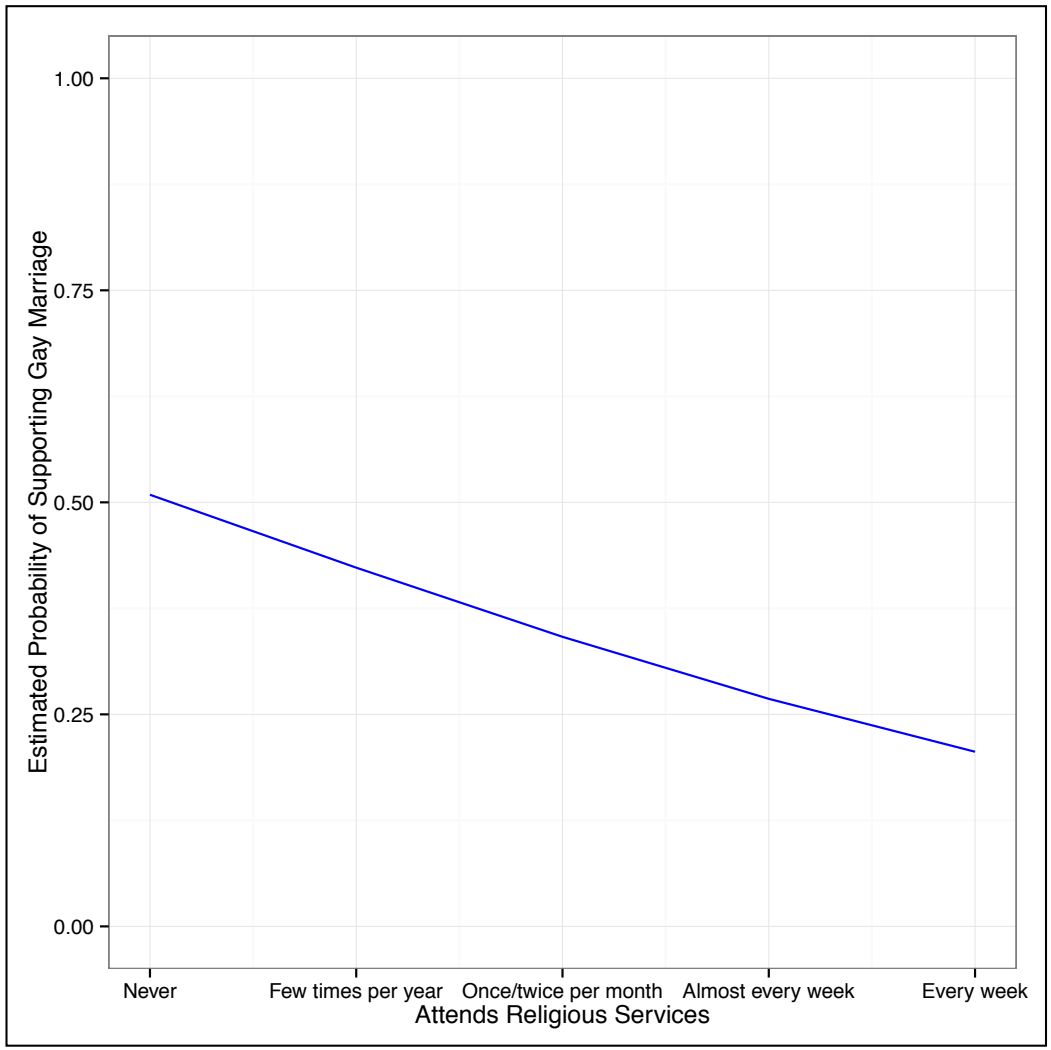
```
> plotdata <- data.frame(
      attend2 = seq(from = 0, to = 4, by = 1)
      )
```

Step 2: Use the `predict()` function to produce a fitted value for each row in the newly created data frame.



Step 3: Plot the fitted values versus the *X*s

```
> ggplot(data = plotdata, aes(x = attend2, y = fitted)) +
     geom_line(color = "blue") +
     scale_x_continuous(
          name = "Attends Religious Services",
          labels = c("Never", "Few times per year", "Once/twice per
                     month", "Almost every week", "Every week")
          ) +
     ylab("Estimated Probability of Supporting Gay Marriage") +
     ylim(c(0, 1)) +
     theme_bw()
```



Interpreting the Plot

- The predicted probability of supporting gay marriage for Americans who never attend religious services is 0.51.
- The effect of attending religious services on the predicted change in probability of supporting gay marriage is non-linear, and depends on how frequently religious services are attended.

Figure 1. Estimated probability of supporting gay marriage as a function of attendance at religious services (N=1,746)

A Shortcut to Plotting the Fitted Model

```
> ggplot(data = grad, aes(x = gre, y = admit)) +
    geom_smooth(method = "glm", family = "binomial(link = "logit"), se = FALSE) +
    ylab("Estimated Probability of Supporting Gay Marriage") +
    scale_x_continuous(
         name = "Attends Religious Services",
         labels = c("Never", "Few times per year", "Once/twice per month", "Almost
                    every week", "Every week")
         ) +
    ylim(c(0, 1)) +
    theme_bw()
```