Assignment #13: GLM

Scenario: Over the past few years, our first-to-second year retention rate has ranged from 77-80%. In other words, 77-80% of our first-year students come back to St. Ambrose for their second year.

Let's see if we can predict whether a student returns or does not return to St. Ambrose for their second year.

2011-13 student data: http://www.bradthiessen.com/html5/data/retention1113.csv
2014 student data: http://www.bradthiessen.com/html5/data/retention14.csv

1. Suppose you were tasked with developing a model to predict whether a student returns to St. Ambrose for their second year. Identify 4 variables you think might predict a student's decision to come back:

   1. _____ 2. _____

   3. _____ 4. _____

2. The dataset for this assignment contains 47 variables measured for 1724 observations (students). The variables include:
   - ACT composite, English, math, reading, and science scores
   - High school GPA and rank; the number of different high schools attended by each student
   - Gender and race
   - The highest level of education completed by each student's mother and father
   - Whether the student is a student athlete
   - The number of credits attempted during each student's first semester; the first semester GPA
   - Each student's major
   - Responses to a survey that measure each student's:
     - commitment to St. Ambrose, academic ability, ability to pay for tuition, satisfaction with St. Ambrose
   - (and other variables)

   In this assignment, I'll need to select a subset of these variables to include in my prediction model. For now, let's use the following logistic regression model:

   $$\ln(\text{odds of returning to SAU}) = b_0 + b_1(\text{ACTcomp}) + b_2(\text{HSgpa}) + b_3(\text{residence}) + b_4(\text{athlete}) + b_5(\text{black})$$

   where ACTcomp = ACT Composite score (ranges from 17 to 35 in this data)
   HSgpa = high school GPA (ranges from 1.00 to 4.00 in this data)
   residence = does a student live on campus (1 = off-campus, 0 = on-campus)
   athlete = (1 = not a student athlete, 0 = student athlete)
   black = (1 = African-American student; 0 = not African-American student)

   I fit this model in R and obtained the output pasted on the next page…

```
Call:
glm(formula = retained ~ hsgpa + act_comp + residence + athlete + black,
                                          family = binomial, data = retention2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3924   0.3154   0.4008   0.5208   2.2767

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.963237   0.543613  -1.772   0.0764 .
hsgpa        0.945571   0.159675   5.922 3.18e-09 ***
act_comp    -0.002036   0.025909  -0.079   0.9374
residence1  -2.937433   0.181208 -16.210  < 2e-16 ***
athlete1     0.318550   0.166907   1.909   0.0563 .
black1      -1.453135   0.310180  -4.685 2.80e-06 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1638.5  on 1586  degrees of freedom
Residual deviance: 1176.7  on 1581  degrees of freedom
  (126 observations deleted due to missingness)
AIC: 1188.7

Number of Fisher Scoring iterations: 5



Confidence Interval bounds:
                   2.5 %      97.5 %
(Intercept) -2.034430752  0.09844430
hsgpa        0.634007788  1.26045090
act_comp    -0.052733403  0.04891276
residence1  -3.299285334 -2.58814752
athlete1    -0.005794063  0.64928171
black1      -2.061672003 -0.84146513



Exponentiated coefficients:
                    OR      2.5 %     97.5 %
(Intercept) 0.38165560 0.13075489 1.10345294
hsgpa       2.57428251 1.88515074 3.52701146
act_comp    0.99796589 0.94863288 1.05012873
residence1  0.05300161 0.03690954 0.07515914
athlete1    1.37513288 0.99422269 1.91416541
black1      0.23383601 0.12724104 0.43107847
```

3. Write out the coefficients for this model:

ln(odds of returning to SAU) = _____ + _____(HSgpa) + _____(ACTcomposite) +

_____(residence) + _____(athlete) + _____(black)

4. Just based on the p-values, which predictors appear to predict whether a student returns to St. Ambrose? Identify whether each variable predicts a higher or lower chance of returning to St. Ambrose. If the predictor is not significant, circle "unknown."

Higher HSgpa predicts a student is……	**MORE**	**LESS**	**UNKNOWN**	likely to return
Higher ACT Composite score predicts a student is….	**MORE**	**LESS**	**UNKNOWN**	likely to return
Living off-campus predicts a student is……..	**MORE**	**LESS**	**UNKNOWN**	likely to return
Student athletes are……	**MORE**	**LESS**	**UNKNOWN**	likely to return
African-American students are……..	**MORE**	**LESS**	**UNKNOWN**	likely to return

5. Suppose we have a student with the following:
   HSgpa = 3.50
   ACTcomposite = 22
   Lives on-campus (residence = 0)
   Is not an athlete (athlete = 1)
   Is not African-American (black = 0)

   What are the log-odds of this student returning to SAU for his or her second year? _____

   What are the odds of this student returning to SAU for his or her second year? _____

   What is the probability this student returns to SAU for his or her second year? _____

6. Suppose we have another student identical to the previous one except for the fact that this student lives off-campus.

   What are the log-odds of this student returning to SAU for his or her second year? _____

   What are the odds of this student returning to SAU for his or her second year? _____

   What is the probability this student returns to SAU for his or her second year? _____

7. Take another look at the exponentiated coefficients of our model.

```
Exponentiated coefficients:
                    OR      2.5 %     97.5 %
(Intercept) 0.38165560 0.13075489 1.10345294
hsgpa       2.57428251 1.88515074 3.52701146
act_comp    0.99796589 0.94863288 1.05012873
residence1  0.05300161 0.03690954 0.07515914
athlete1    1.37513288 0.99422269 1.91416541
black1      0.23383601 0.12724104 0.43107847
```

Interpret the OR value for the variable black.

0.2338 represents: _____

_____

8. I had R add our predictor variables sequentially and test the contribution of each predictor to our overall prediction of student retention:

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: retained
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      1586     1638.5
hsgpa      1   109.39      1585     1529.1 < 2.2e-16 ***
act_comp   1     4.14      1584     1525.0   0.04199 *
residence  1   325.18      1583     1199.8 < 2.2e-16 ***
athlete    1     2.20      1582     1197.6   0.13788
black      1    20.92      1581     1176.7 4.801e-06 ***
```

Based on this output, what can we conclude about the value of ACTcomposite as a predictor of student retention?

_____

_____

9. I then fit the following simplified model:

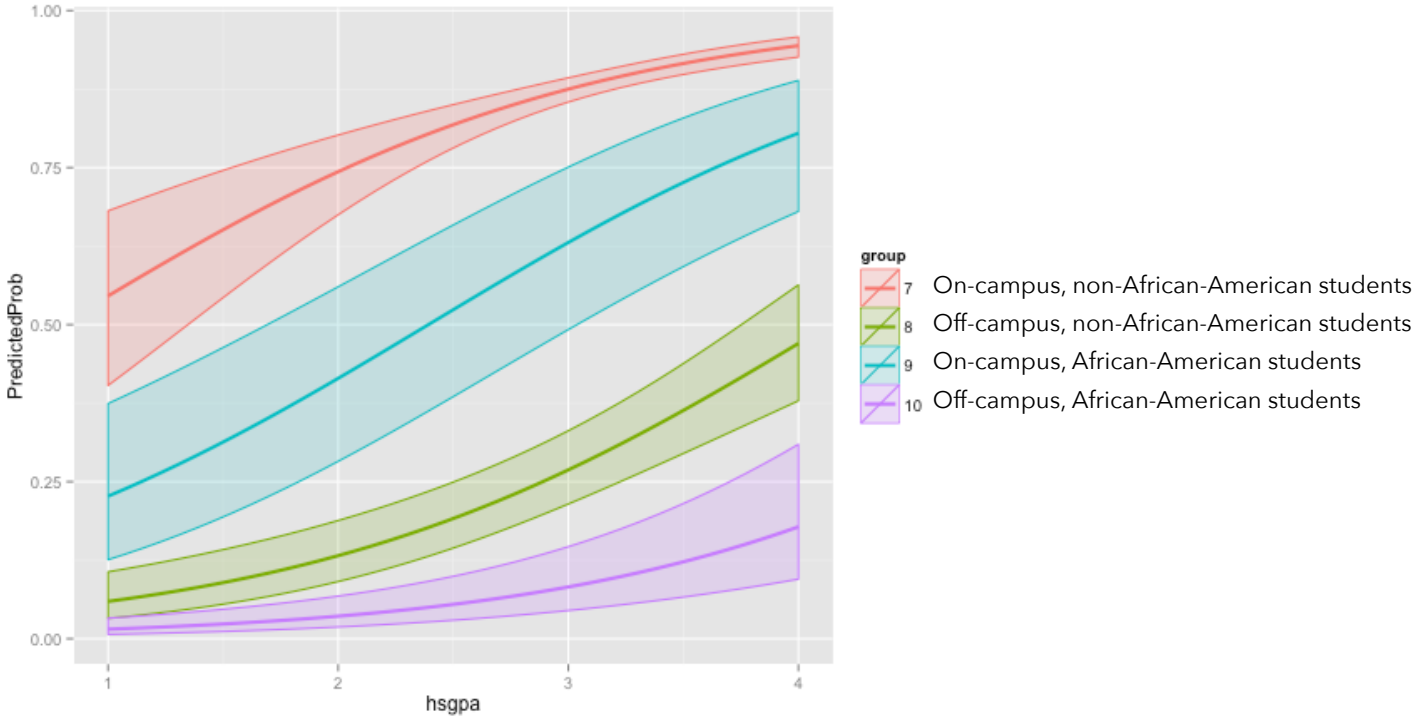
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6978     0.4240  -1.646   0.0998 .
hsgpa         0.8808     0.1348   6.537 6.28e-11 ***
residence1   -2.9458     0.1744 -16.891  < 2e-16 ***
black1       -1.4094     0.3026  -4.658 3.20e-06 ***
```

(continued on next page)

Here are the exponentiated coefficients of this model:

```
                    OR      2.5 %    97.5 %
(Intercept) 0.49769142 0.21710463 1.1457026
hsgpa       2.41292436 1.85580040 3.1487196
residence1  0.05255923 0.03711802 0.0735883
black1      0.24429733 0.13498105 0.4438793
```

And here's a plot of the model:



Finally, I used this model to make predictions for our Fall 2014 students. Based on this model, the average retention rate for this class is predicted to be 84% (with a confidence interval between 82.5% and 85.5%).

Scenario: As you know, the Titanic sank in the Atlantic ocean during its maiden voyage from the UK to New York after colliding with an iceberg.

The dataset we're going to investigate has the following records for each of the 1309 passengers on the Titanic:

- survival = Survival (0 = No; 1 = Yes)
- pclass = Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- name = Name
- sex = Sex
- age = Age
- sibsp = Number of Siblings/Spouses Aboard
- parch = Number of Parents/Children Aboard
- ticket = Ticket Number
- fare = Passenger Fare
- cabin = Cabin
- embarked = Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Of the 1309 individuals in this dataset, 500 (38.2%) survived. We're going to construct some models to predict who survived.

To help ensure we don't overfit our data, let's first split this dataset into two pieces:

Training data = 900 randomly selected rows from our dataset
Test data = the remaining 409 unselected rows from our dataset

We'll fit our models to the training data and then see how well they predict survival in the test dataset.

Titanic data: http://www.bradthiessen.com/html5/data/titanic.csv
R code for this entire assignment: **http://www.bradthiessen.com/html5/data/titanic.csv**

10. "Women and children first!" is a statement you would never hear me say on a boat that is sinking. But I have heard this phrase, so let's see if age and gender give us a good prediction.

I fit the model **ln(odds of surviving) = b_0 + b_1(gender) + b_2(age)** and obtained the following output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.953783   0.239598   3.981 6.87e-05 ***
sexmale     -2.396153   0.182771 -13.110  < 2e-16 ***
age          0.001847   0.006533   0.283    0.777

    Null deviance: 956.05  on 711  degrees of freedom
Residual deviance: 754.43  on 709  degrees of freedom
  (188 observations deleted due to missingness)
AIC: 760.43


Exponentiated coefficients
                    OR      2.5 %    97.5 %
(Intercept) 2.59551083 1.63391405 4.1855733
sexmale     0.09106758 0.06327271 0.1296023
age         1.00184876 0.98906771 1.0147600
```

11. Calculate the probability of a female child (age = 8) surviving. Then, calculate the probability of a male adult (age = 40) surviving. Calculate the ratio of these two values to find the relative probability.

P(survive | female child) = _____

P(survive | male adult) = _____

Relative probability = _____

12. In the middle of the output, there's a bold line that states: "(188 observations deleted due to missingness)." Those values were deleted because they represent individuals with no recorded age. Let's assume an individual with a missing age is an adult. With this assumption, we can recode our data with this logic:

Create a new variable named "child"
If age < 18, then child == 1
Otherwise, child == 0

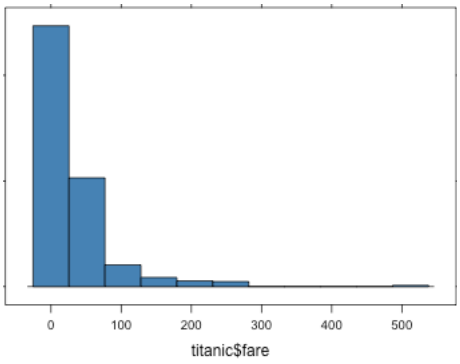
This removes much of the information about the age of the passengers, but it does fill-in those missing values. Let's fit a model with this new variable (replacing age): **ln(odds of surviving) = b_0 + b_1(gender) + b_2(child)**

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8836     0.1275   6.928 4.25e-12 ***
sexmale      -2.3643     0.1632 -14.484  < 2e-16 ***
child         0.2622     0.2645   0.991    0.322

    Null deviance: 1190.30  on 899  degrees of freedom
Residual deviance:  944.06  on 897  degrees of freedom
AIC: 950.06

Exponentiated coefficients
                    OR      2.5 %    97.5 %
(Intercept) 2.59551083 1.89177584 3.1207340
sexmale     0.09106758 0.06796806 0.1289328
age         1.00184876 0.77111942 2.1786271
```

Repeat the calculations you did above to find the relative probability of surviving for a female child and adult male. Use the coefficients from this new model.

Relative probability = _____

13. That last model took care of our missing data problem, but I bet we could improve our model by including another variable. Rich people were probably more likely to survive, so let's use the "fare" variable as a proxy for economic status.



The histogram to the right shows the fares paid by each passenger. I know the Titanic crashed in 1912, so I can use an inflation calculator to see how much these tickets would cost in 2014:

http://data.bls.gov/cgi-bin/cpicalc.pl?cost1=10&year1=1913&year2=2014

This calculator tells me that $10 in 1912 is equivalent to $240 in 2014. That's

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.21880    0.18059   1.212    0.226
sexmale     -2.23531    0.16971 -13.171  < 2e-16 ***
child1       0.07825    0.26564   0.295    0.768
Fare210-20   0.36086    0.23792   1.517    0.129
Fare220-30   0.60701    0.25189   2.410    0.016 *
Fare230+     1.47663    0.21079   7.005 2.46e-12 ***


    Null deviance: 1190.30  on 899  degrees of freedom
Residual deviance:  890.45  on 894  degrees of freedom
AIC: 902.45

Exponentiated coefficients
                  OR      2.5 %    97.5 %
(Intercept) 1.2445770 0.87363287 1.7755003
sexmale     0.1069594 0.07634206 0.1485584
child1      1.0813961 0.64016684 1.8172592
Fare210-20  1.4345689 0.89754770 2.2837948
Fare220-30  1.8349343 1.11767495 3.0041567
Fare230+    4.3781514 2.90614333 6.6458596
```

14. Let's add one more variable to our model. If you were to look at the names of the individuals in this dataset, you would find names like:
    - Lesurer, Mr. Gustave J
    - Duff Gordon, Lady.
    - Rothes, the Countess. of (Lucy Noel Martha Dyer-Edwards)

    Those titles (e.g., Mr., Lady, the Countess) might give us more information about the economic status of these individuals. Using R, I separated the titles from each person's name and found the following frequencies in our training dataset:

```
  Capt     Col     Don    Dona      Dr Jonkheer    Lady
     1       4       1       1       8        1       1
 Major  Master    Miss    Mlle    Mme      Mr     Mrs
     2      61     260       2       1     757     197
    Ms     Rev     Sir the Countess
     2       8       1       1
```

I'm going to combine some of these categories:

Mlle = MMe (Madame) + Mlle (Mademoiselle)
Sir = Capt + Don + Major + Sir
Lady = Dona + Lady + the Countess + Jonkheer (Dutch royalty)

Then, looking at the original dataset, I see two variables I haven't yet used:

- sibsp = Number of Siblings/Spouses Aboard
- parch = Number of Parents/Children Aboard

I'm not sure I know why these variables would impact the chances of survival, but let's combine them and throw them into our prediction model. I'll add them together to create a variable named "familysize."

I can now fit this model: **ln(odds of surviving) = b_0 + b_1(gender) + b_2(child) + b_3(fare) + b_4(title) + b_5(familysize)**

```
Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.72804    0.42582   6.407 1.49e-10 ***
sexmale       -2.58574    1.50178  -1.722  0.08511 .
child1         0.19334    0.33334   0.580  0.56192
Fare210-20     0.38852    0.25636   1.516  0.12964
Fare220-30     0.92541    0.28051   3.299  0.00097 ***
Fare230+       2.07975    0.25536   8.144 3.81e-16 ***
titleDr        0.56318    1.81996   0.309  0.75698
titleLady     12.60586  833.92369   0.015  0.98794
titleMaster    2.03887    1.56605   1.302  0.19295
titleMiss     -0.71160    0.29961  -2.375  0.01754 *
titleMlle     12.40316 1029.12152   0.012  0.99038
titleMr       -0.62362    1.48428  -0.420  0.67438
titleMrs            NA         NA      NA       NA
titleRev     -14.75145  633.32473  -0.023  0.98142
titleSir       0.72010    1.93567   0.372  0.70988
familysize    -0.54829    0.08104  -6.766 1.33e-11 ***

    Null deviance: 1190.30  on 899  degrees of freedom
Residual deviance:  812.27  on 885  degrees of freedom
AIC: 842.27
```

```
Exponentiated coefficients
                     OR         2.5 %        97.5 %
(Intercept) 1.530285e+01  6.808592e+00  3.625575e+01
sexmale     7.534033e-02  2.655653e-03  2.134839e+00
child1      1.213291e+00  6.302789e-01  2.335132e+00
Fare210-20  1.474792e+00  8.896165e-01  2.434150e+00
Fare220-30  2.522913e+00  1.453715e+00  4.372898e+00
Fare230+    8.002461e+00  4.885165e+00  1.331014e+01
titleDr     1.756248e+00  3.780773e-02  8.077053e+01
titleLady   2.983028e+05  2.018948e-42            NA
titleMaster 7.681923e+00  2.479033e-01  2.388904e+02
titleMiss   4.908576e-01  2.703222e-01  8.772456e-01
titleMlle   2.435704e+05  2.033055e-65            NA
titleMr     5.360001e-01  1.933602e-02  1.482430e+01
titleMrs              NA            NA            NA
titleRev    3.922154e-07 1.843144e-124 1.122984e-181
titleSir    2.054632e+00  3.921651e-02  1.359976e+02
familysize  5.779351e-01  4.900285e-01  6.737272e-01
```

15. Now that we have four different models, let's see how well they predict survival on our test dataset. Remember, our test dataset includes 409 observations that were not used to estimate the coefficients of our models.

Observed results in the test data set: 563 died; 337 survived

When I fit these models, I get a predicted probability of each individual surviving or dying. To compare these with the actual data, I'm going to use the following rule:

If predicted probability of survival > 0.50, then we predict the individual to survive
If predicted probability of survival < 0.50, then we predict the individual to die

Let's see how our predictions hold on the test data:

Model #1: survival = f(sex, age)

	Observed deaths	Observed survivors
Predicted deaths	220	63
Predicted survivors	26	100

Model #2: survival = f(sex, child)

	Observed deaths	Observed survivors
Predicted deaths	210	51
Predicted survivors	36	112

Model #3: survival = f(sex, child, fare)

	Observed deaths	Observed survivors
Predicted deaths	210	51
Predicted survivors	36	112

Model #4: survival = f(sex, child, fare, family size) – (note: I had some problems with the title variable, so I removed it)

	Observed deaths	Observed survivors
Predicted deaths	199	52
Predicted survivors	47	111

On the next page, we'll calculate an index to see how well each model fit.

16. We can use a statistic called *Cohen's Kappa* to measure the agreement between two measures. Given the following table:

	Observed deaths	Observed survivors	Total
Predicted deaths	a	b	W
Predicted survivors	c	d	X
Total	Y	Z	N

We'd calculate Kappa as $\kappa = \dfrac{(a+d) - \left(\dfrac{WY}{N} + \dfrac{XZ}{N}\right)}{1 - \left(\dfrac{WY}{N} + \dfrac{XZ}{N}\right)}$

Use the online applet at http://www.graphpad.com/quickcalcs/kappa1/ to calculate Kappa for each table on the previous page:

Model 1 kappa = _____

Model 2-3 kappa = _____

Model 4 kappa = _____

Perfect agreement would yield Kappa = 1. Anything less than 1.0 represents less-than-perfect agreement. If Kappa is negative, it means the model agreed less with the actual data than we'd expect just by chance.

Which model provided us the best prediction? Model # _____.

17. Since roughly 38% of the passengers survived, maybe we should only predict survival for the individuals with predicted probabilities of surviving of 38% or greater. If we do that, we get the following results for models 3-4:

Model #3: survival = f(sex, child, fare)

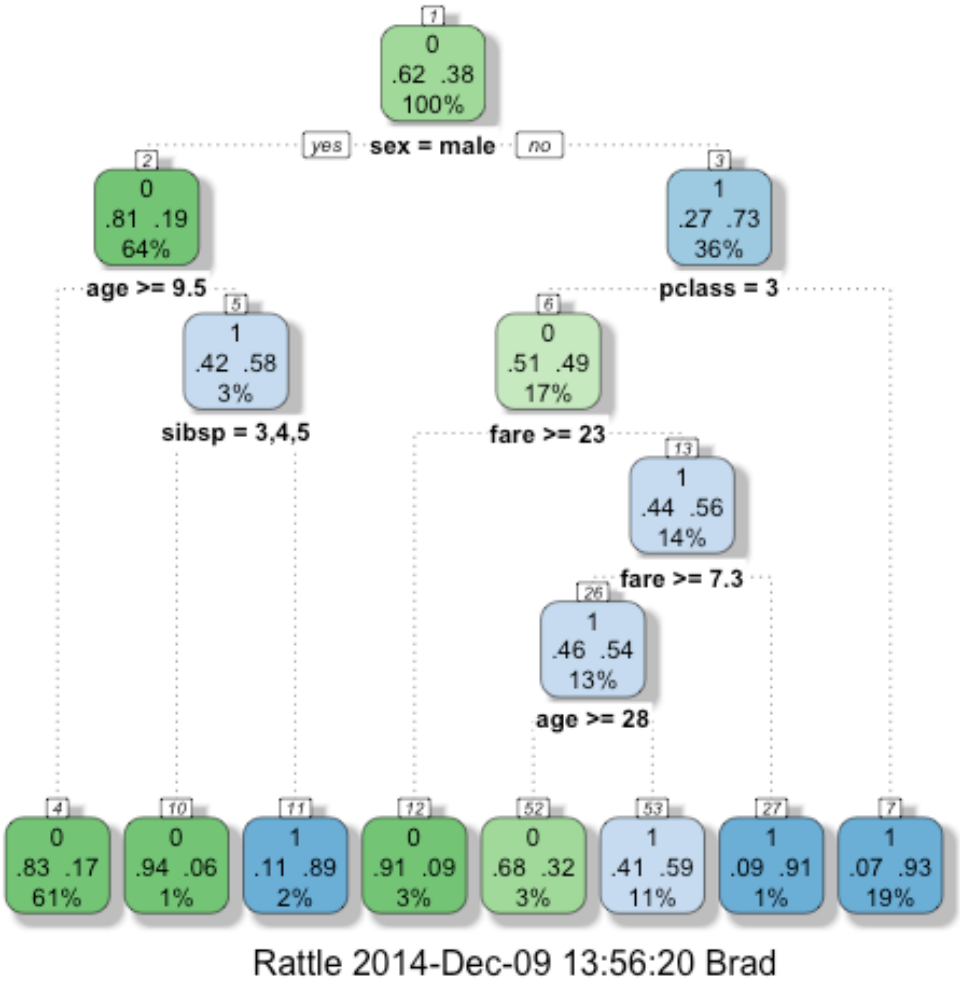
	Observed deaths	Observed survivors	
Predicted deaths	205	47	Kappa = _____
Predicted survivors	41	116	

Model #4: survival = f(sex, child, fare, family size)

	Observed deaths	Observed survivors	
Predicted deaths	182	43	Kappa = _____
Predicted survivors	64	120	

Calculate Kappa for each model.

18. If we have time, I'll show you how to construct decision trees to make predictions. Here's the results of a decision tree for this Titanic dataset:



Reading these trees can be tricky, so let's give it a shot.