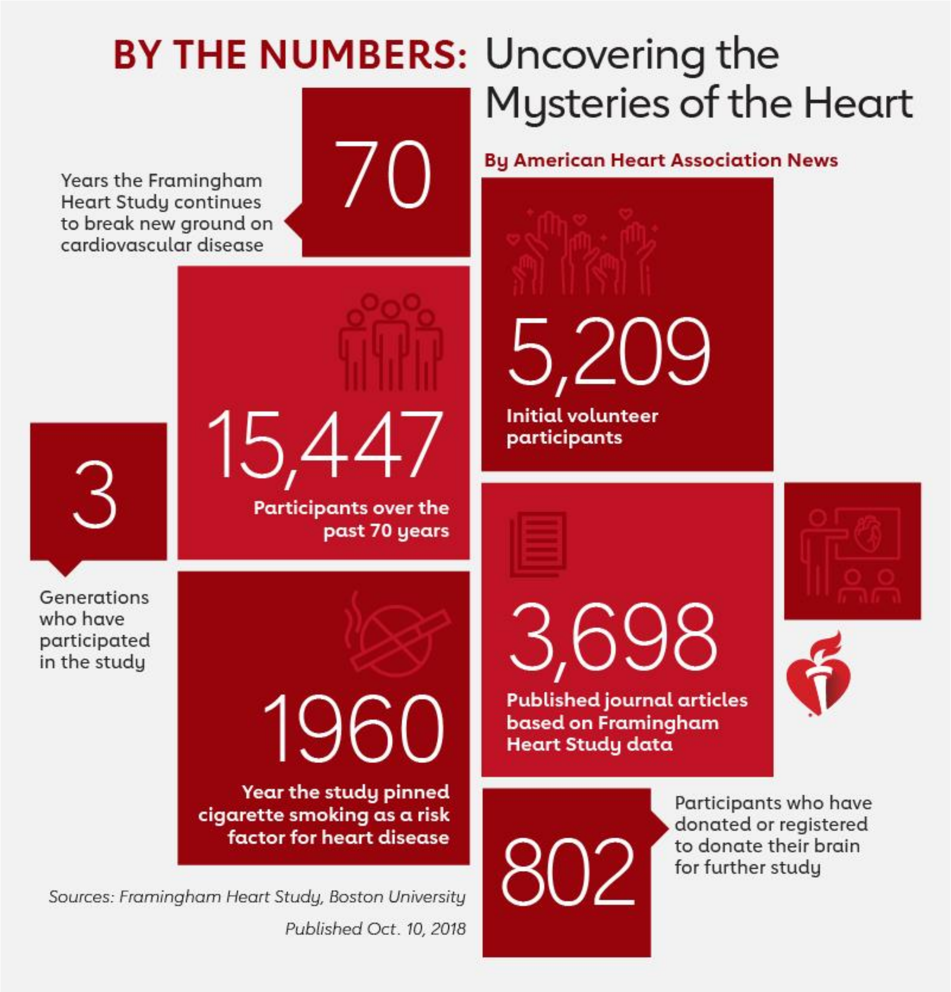
Lab 3

The Framingham Heart Study is a long term prospective study of cardiovascular disease among a population of subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects over the course of three generations. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes.

You will find the data file ***framinghamHeart.csv***, which you can load as `dff`. This is a subset of the data collected as part of the Framingham study. Participant clinic data was collected during three examination periods, approximately 6 years apart, from roughly 1956 to 1968. Each participant was followed for a total of 24 years for the outcome of a specified set of adverse health events. The dependent variable is **TenYearCHD**, specifying whether a subset of events associated with chronic heart disease occurred within 10 years of follow up. The variables are defined below. The purpose of the study is to determine the risk factors of heart disease.


BY THE NUMBERS: Uncovering the Mysteries of the Heart
By American Heart Association News
70
Years the Framingham Heart Study continues to break new ground on cardiovascular disease
5,209
Initial volunteer participants
15,447
Participants over the past 70 years
3
Generations who have participated in the study
3,698
Published journal articles based on Framingham Heart Study data
1960
Year the study pinned cigarette smoking as a risk factor for heart disease
802
Participants who have donated or registered to donate their brain for further study
Sources: Framingham Heart Study, Boston University
Published Oct. 10, 2018

Data Dictionary

Variable	**Description**	**Coding**
gender	Male or Female	0 = Female; 1 = Male
age	Age of the patient	
education	Highest level of education achieved	1 = High School; 2 = High School Diploma or GED; 3 = Some college or vocational School; 4 = College degree
currentSmoker	Indicates if the person is currently a smoker or not	0 = Not a smoker; 1 = Is a smoker
cigsPerDay	The number of cigarettes the person smoked on average in one day	
BPMeds	Whether the patient was on blood pressure medication	0 = Not on BP meds; 1 = On BP meds
prevalentStroke	Whether the patient previously had a stroke	0 = Free of disease; 1 = Stroke
prevalentHyp	Whether the patient has hypertension (high blood pressure)	0 = Free of disease; 1 = Hypertension
diabetes	Whether the patient has diabetes	0 = Free of disease; 1 = Diabetes
totChol	Total cholesterol level	mg/dL
sysBP	Systolic blood pressure	mmHg
diaBP	Diastolic blood pressure	mmHg
BMI	Body Mass Index	Weight (kg) / Height (meter-squared)
heartRate	Heart rate	Beats/Min (Ventricular)
glucose	Glucose level	mg/dL
TenYearCHD	Coronary heart disease	‘0' indicates the event did not occur during the 10-year follow

		up, and ‘1' indicates an event did occur during the follow up

Data Analysis

Before you start, **load the “caret” library** in addition to the usual four libraries we always load.

In addition, pay attention to what R reports after you load the dataset:

```
Parsed with column specification:
cols(
  gender = col_double(),
  age = col_double(),
  education = col_double(),
  currentSmoker = col_double(),
  cigsPerDay = col_double(),
  BPMeds = col_double(),
  prevalentStroke = col_double(),
  prevalentHyp = col_double(),
  diabetes = col_double(),
  totChol = col_double(),
  sysBP = col_double(),
  diaBP = col_double(),
  BMI = col_double(),
  heartRate = col_double(),
  glucose = col_double(),
  TenYearCHD = col_double()
)
```

Notice that R reads all the columns as numbers. You know from the data dictionary that some variables are supposed to be factors. You need to ask R to convert them into factors:

i. Create a list of columns that are supposed to be factors:

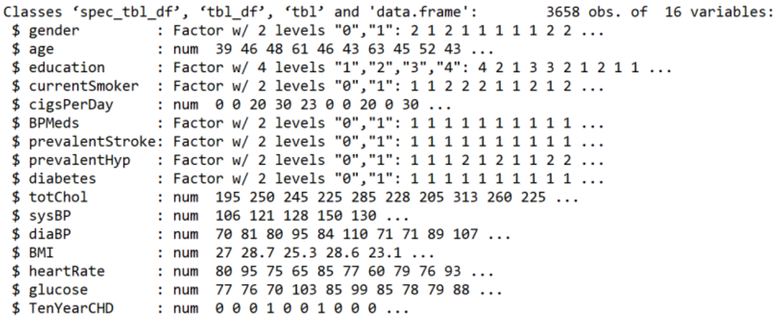
```
colsToFactor <- c('gender', 'education', 'currentSmoker', 'BPMeds', 
'prevalentStroke', 'prevalentHyp', 'diabetes')
```

ii. Ask R to replace (overwrite) selected variables with their factor conversions:

```
dff <- dff %>%
  mutate_at(colsToFactor, ~factor(.))
```

=> What do you think ***mutate_at*** *does?*

Now, if you run `str(dff)`, you will see that the variables in your data are correctly identified:



```
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame':       3658 obs. of  16 variables:
 $ gender         : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 2 2 ...
 $ age            : num  39 46 48 61 46 43 63 45 52 43 ...
 $ education      : Factor w/ 4 levels "1","2","3","4": 4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker  : Factor w/ 2 levels "0","1": 1 1 2 2 2 1 1 2 1 2 ...
 $ cigsPerDay     : num  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ prevalentStroke: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ prevalentHyp   : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 1 2 2 ...
 $ diabetes       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ totChol        : num  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : num  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : num  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD     : num  0 0 0 1 0 0 1 0 0 0 ...
```

1. **Data exploration:** To explore visually whether blood pressure levels and total cholesterol levels are associated with heart disease, create boxplots of *sysBP*, *diaBP*, and *totChol*, broken up by the levels of *TenYearCHD*. [**Hint:** Dynamic plots may help understanding!]

2. **Data preprocessing:**

(i) Read the data file into R. Set the seed to **123** and split the data into `dffTrain` and `dffTest`. Randomly sample 70% of the data for training, and use the rest as test dataset.

(ii) What are the proportions by gender in your training vs. test set? How does the distribution of age look? Looking at these, do you observe any signs of a sampling bias?

Training set – 55:45

Test set – 56:44

Maximum people are in the 40-50 years age group, while minimum are in the 60-70 years old range. The distribution is slightly skewed towards the right suggesting that there is more data about younger people than older people. The graph also shows spikes at certain ages like near age 40, 48 etc. These lead us to believe that the sample space is not equally distributed and can introduce sample bias.

Hints:

[A] It's time to use R like a pro! You can pipe your *dffTrain* into the `group_by(`*`variable`*`)` function and then into **`tally()`** *-no arguments-* to get the counts across a group.

- ❑ To add percentages, pipe one more step into `mutate(pct = 100*n/sum(n))`

[B] For a continuous variable like age, there are so many groups, right? Each age is practically a different group. In such cases, you may want to create your own groups.

- ❑ You can use *`ageGroup=cut_interval(age, length=10)`* in group_by()

[C] You can also create a histogram for age, which probably makes more sense.

❑ After creating the histogram, try adding `fill=gender` into aes() of ggplot(), and see what happens. In addition, define `color='black'` inside the histogram!

3. **Linear probability model:** Build a linear probability model `fitLPM` using all variables in `dffTrain`. Make sure to check for collinearity[1] by both thinking about the variables, and using VIF values as guiding signals, and take necessary precautions. You know how to mitigate collinearity (if not, please ask during the lab!). After finalizing the model, which of the variables are statistically significant at the 95% level? What does this model tell you about the risk factors of heart disease? Do you have any reservations? Discuss.

Following are the variables that are statistically significant at the 95% level –

gender1, cigsPerDay, prevalentStroke1, prevalentHyp1, sysBP, heartrate, glucose.

The model tells us that males are more susceptible to heart diseases. Smoking cigarettes, having had strokes in the past or having high blood pressure also increases the chance of having heart diseases. Higher the systolic blood pressure, more the chance of getting a heart disease. Higher heart rate corresponds to low chance of getting a heart disease while high glucose level corresponds to increased chance of heart disease.

While rest seems plausible, high heart rate should correspond to higher risk of heart diseases which is not the case according to the model. Additionally, total cholesterol is not a significant statistic according to the model which is contrary to the fact that high cholesterol may lead to heart disease. This shows that the model might not be entirely accurate at judging the dependent variable.

Hints:

[A] To include all the variables, use a full stop **.** To exclude a variable, use a negative **-**

[B] Run diagnostics to see whether this model violates the linear regression assumptions.

[1] Likely multicollinearity. If "multicollinearity problem is extreme: any variable in the model can be written as a linear combination of all of the other variables in the model. Essentially, this means that we can never know exactly which variables (if any) truly are predictive of the outcome, and we can never identify the best coefficients for use in the regression. At most, we can hope to assign large regression coefficients to variables that are correlated with the variables that truly are predictive of the outcome." ISLR p. 243

4. Speaking of using R like a pro, a better way to run a model and create a results table with predictions is as follows. Please run this code to make predictions using the LPM model and store them into *resultsLPM*[2]

```
resultsLPM <-
   lm( …fill in here… ) %>%
   predict( …fill in here… ) %>%     => Use the option type='response' for probabilities
   bind_cols(dffTest, predictedProb=.) %>%      => The dot marks where to pipe into
   mutate(predictedClass = …fill in here… )     => Use 50% as cutoff for classification
```

Inspect `resultsLPM`. Then, **copy and paste your code from Q2-ii** and check the prevalence of *TenYearCHD* in the *test dataset* this time. How many people have heart disease in reality (in the test dataset)? Run the same code for *predictedClass* in the *resultsLPM*. How many people did the model predict having heart disease? Compare and report your observations.

In reality, 172 people have heart disease. The model predicted 10 people to have heart disease. The difference in number may be because of classification errors in the model. Many people with heart disease were classified as those not having heart disease by the model.

Before you continue:

You may have noticed that we did not convert TenYearCHD into a factor yet, even though it is a factor. This is because we wanted to use it in a linear model. It is time to make it a factor.

❑ Use `mutate()` to convert TenYearCHD to a factor both in *dffTrain* and *dffTest* datasets.

[2] You can replicate this idea for any other model to make predictions -including the ones you did last week. When you are using this chunk of code for a linear regression, you don't need the last line because you don't need a conversion into classes. Instead, I would change *bind_cols(dffTest, **predictedProb**=.)* into *bind_cols(dffTest, **predictedValues**=.)* for a better understanding in a linear model.

5. **Logistic regression:** Build a logistic regression using the predictor variables you decided to keep in the model you built in Q3. Which variables are statistically significant at the 95% level? Compare your results with the results you obtained from the model in Q3.

   Hint: See the appendix for an annotated logistic regression output in R with the definitions.

   Following variables are statistically significant at 95% -

   gender1, age, cigsperday, prevalentHyp1, totChol, sysBP, hearRate, glucose

   While the earlier model does not consider age and total cholesterol level as statistically significant for determining heart disease, this model does. On the other hand, this model does not consider the statistical significance of a person having a history of stroke on a probable heart disease, whereas the earlier model does. High residual deviance and high AIC suggests that the model has more independent variables than actually required.

   Interpret the following variables: *age*, *gender*, and *diabetes* (whether significant or not):

   ❑ **Hint:** You can run **`exp(coef(fit))`** after a logistic model to exponentiate the coefficients of all variables at once, and use them in your interpretations.

   ❑ **Type these interpretations AFTER completing the lab unless you have any questions.**

   Age: On an average, increase by 1 year in age is associated with 6% rise in risk of getting heart disease, keeping all other variables constant. This is statistically significant since the p-value is low.

   Gender: On an average, male have a 52% more chance of getting a heart disease than female, keeping all other variables constant. This is also statistically significant as its p-value is low.

   Diabetes: On an average, people with diabetes have a 0.5% less chance of having a heart disease, as compared to those who do not have diabetes, keeping all other variables constant. This variable is not statistically significant as the p-value is not low.

6. Create a new results table ***resultsLog*** by using the logistic model. Let's continue like a pro.

   Hint: You will follow the same steps you took in Q4 but this time for logistic regression. This means, **your predictedClass will need to be defined as a factor** (you know how to do this!).

How many people did the logistic model predict having heart disease? Report your observations and compare them with the actual values, and the predictions of the linear probability model from Q4. Do you think the logistic model is an improvement? Why?

According to this model 19 people have heart disease. It is still very less than the actual value of 172 which may again be possible because of classification errors. However, this model is an improvement over the linear probability model as the numbers have increased.

Hint: For now, continue to use your code from Q2-ii to create the tables for comparison.

7. It is time to create a confusion matrix, a final step before evaluating performance (which we will cover next week). As you're using R like a pro, it is so easy to create a confusion matrix.

- ❑ Pipe the *resultsLog* dataframe you created in **Q6** into the function `conf_mat(truth = ..., estimate = ...)`
- ❑ **Optional:** Pipe one more step into `autoplot(type = 'heatmap')` to color code. This is useful when more than two classes are involved. For now, this is just a learning point.

Explain what the matrix tells you in addition to what you learned from the tables in **Q6**.

The confusion matrix tells us that the model correctly predicted 605 patients that do not have heart disease and 8 patients that have heart disease. On the other hand it falsely predicted that 474 patients do not have heart disease, who actually have heart disease (false negative) and 11 patients have heart disease, who actually don't have heart disease (false positive).

8. No analysis is complete without a visualization. Plot the relationship between the statistically significant variables (*age*, *cigsPerDay*, *totChol*, *glucose*) and the probability of heart disease:

- ❑ Note that you stored the predicted probabilities as *predictedProb* in the *resultsLog* in Q6.
- ❑ Use `geom_point()` and `geom_smooth()` after `ggplot()`, without adding any parameters
- ❑ Be creative. For example, add `color=currentSmoker` (or `=gender`) into the `aes()`
- ❑ Add a title for the plots, and label both axes [**Hint:** You can use the labs() function]

Discuss your observations.

We can observe from the first plot that as age increases the probability of getting heart disease increases. This probability is more for a male than for a female. Additionally, the chance of getting is the highest for a male who smokes. It is the lowest for a female who doesn't smoke.

From the second plot we see that the probability of getting heart disease increases for those who smoke more cigarettes, in the case of male. We can't say anything definitive in the case of female as the trend is not clear. Overall, the probability of getting a heart disease by a male smoking a particular amount of cigarettes per day is higher than a female, smoking the same amount of cigarettes per day.

Third graph shows us that as the cholesterol level increases in patients the risk of getting a heart disease increases. Also, there is very little difference between the people who smoke and those who don't, when measuring the risk of getting heart disease by high cholesterol level. This suggests that smoking has very little impact on the cholesterol level of people.

The fourth graph shows that as the glucose level increases in a patient, the risk of getting a heart disease increases. Interestingly, here women are more at risk of getting a heart disease via high glucose level than men. Generally, male smokers here are more prone to heart disease while female non-smokers are more susceptible to heart diseases sourcing from high glucose levels.

Switching to a new framework "Caret" we will continue to use in this course from now on:

9. You already loaded the "caret" library at the beginning. If not, load it now. Replicate the analysis in Question 6, this time using the caret library. Use Appendix II[3] for guidance.

❑ Name the results table resultsLogCaret and create it using the train function.
❑ Inspect resultsLogCaret carefully, compare it with resultsLog from Q6 and discuss.
❑ Create the confusion matrix using caret, and compare it with the one in Q7. Discuss.
❑ Don't worry about the rest of the output after the matrix. We will discuss it next week!

resultsLogCaret and resultsLog are the same as both the models are the same rendered by just different methods. This analysis has more accuracy than that in Q7 as it has high number of predictions for people with heart disease who actually have heart disease and those not having heart disease who actually don't have heart disease. Further the analysis shows that it falsely predicted that only 160 people do not have heart when they had a heart disease and only 6 people have heart disease when they actually don't have heart disease. These numbers are very low as compared to those in Q7 suggesting that this analysis is better at predicting the disease.

[3] If you made it to this point, ask me for the handout that includes Appendix II and III.

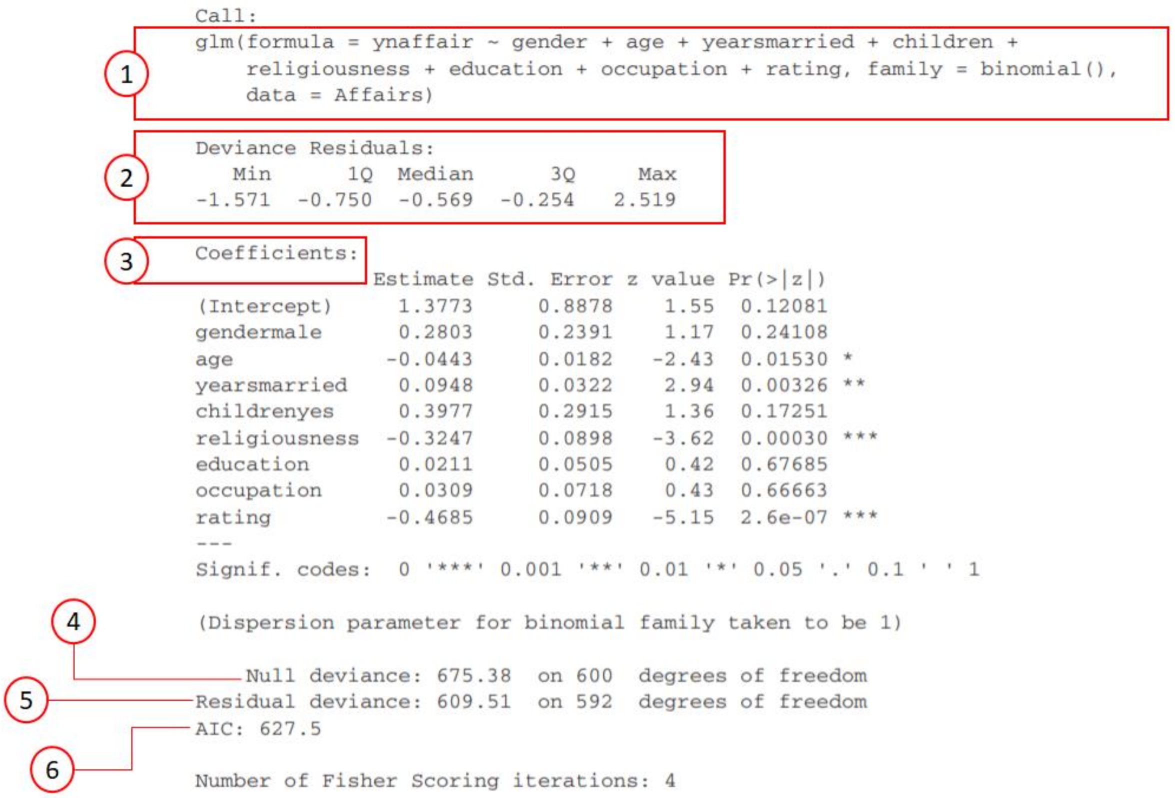
10. Now that you have learned how to use logistic regression for classification, and how to do so **using the caret library**, you can solve another business problem for *Banco Portugal*. See Appendix III for the details of the dataset. The bank runs a telemarketing campaign for a savings account. Have you ever received one of those promotions by the way? “Open a savings account today and get XXX$ bonus!” See this month’s promotions by clicking here.

    Banco Portugal hires you to predict whether a customer will open an account. The bank will use your model to develop promotional campaigns with higher conversion rates. Load the data, make conversions of variables as you see fit, and build logistic regression models using the caret library. Explore at least three alternative models, compare their performance, and pick a final model. Show your full work in the R Notebook. Below, discuss only your findings, your final decision, and explain how your final model helps Banco Portugal with its purpose.

    The best model was achieved by removing newcustomer, agegroup, marital and education from the set of independent variables. An accuracy of 0.903 was attained which is better than when all the variables are included. So I would recommend Banco Portugal to not be bothered with a customer being new or not while running a campaign. Age group was a redundant variable as age is already present in the model. Marital status has no significance on whether an individual opens a bank account or not. Similarly a person’s education holds no significance. Instead of focusing on these details of a customer, the bank can focus on things like whether the person has a personal or housing loan, how long ago was the person last contacted, whether the person has credit in default or not etc.

Appendix I: How to run logistic regression in R and read the regression output

The output from summary() may seem overwhelming at first, so let's break it down one item at a time:



```
Call:
glm(formula = ynaffair ~ gender + age + yearsmarried + children +
    religiousness + education + occupation + rating, family = binomial(),
    data = Affairs)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.571   -0.750   -0.569   -0.254    2.519

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.3773     0.8878    1.55  0.12081
gendermale      0.2803     0.2391    1.17  0.24108
age            -0.0443     0.0182   -2.43  0.01530 *
yearsmarried    0.0948     0.0322    2.94  0.00326 **
childrenyes     0.3977     0.2915    1.36  0.17251
religiousness  -0.3247     0.0898   -3.62  0.00030 ***
education       0.0211     0.0505    0.42  0.67685
occupation      0.0309     0.0718    0.43  0.66663
rating         -0.4685     0.0909   -5.15  2.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 675.38  on 600  degrees of freedom
Residual deviance: 609.51  on 592  degrees of freedom
AIC: 627.5

Number of Fisher Scoring iterations: 4
```

#	Item	Description
1	Formula	Like it was in the linear regression, the *glm()* formula describes the relationship between the dependent and independent variables. Note that you need to include *family = 'binomial'* as an argument.
2	Deviance Residuals	Because the difference between the observed and the fitted values are not very informative in a logistic regression, R reports the deviance residuals, which are the signed square roots of the ith observation to the overall deviance, calculated as follows: $d_i = \text{sgn}(y_i - \hat{y}_i)\left\{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}^{(1/2)}$

3	Coefficients	The regression coefficients show the change in log(odds) in the dependent variable for a unit change in the predictor variable, holding all other predictor variables constant. Because log(odds) are difficult to interpret, we usually exponentiate the coefficients and convert them into the odds scale: exp(the coefficient of yearsmarried) = exp(0.0948) = 1.10, which means a 1-year increase in the number of years married is associated with an increase in the odds of an affair by a factor of 1.10 (about a 10% increase), holding everything else constant. *What about a 10-year increase in the number of years married?* If you interpret a categorical variable like gendermale, exp(0.2803)=1.32 becomes the odds ratio. Therefore, the odds of a male having an affair are about 32% higher than the odds of a female doing so, holding everything else constant. You can exponentiate all coefficients by running `exp(coef(fit))`
4-5	Null Deviance, and Residual Deviance	The *null deviance* shows how well the dependent variable is explained by a model that includes only the intercept. The *residual deviance* shows how well the dependent variable is explained by a model that includes all the independent variables.
6	AIC	The Akaike Information Criterion (AIC) provides a method for assessing the quality of your model through comparison of related models. It's based on the Deviance measure, but includes a penalty for including additional independent variables. Much like adjusted R-squared, it intends to help you leave irrelevant predictors out. However, unlike adjusted R-squared, the reported number itself is not meaningful. When you compare nested models[4], you should select the model that has the smallest AIC. For BIC, run `BIC(fit)` after a regression, where *fit* is the model name, and R will report the BIC score. All of this also applies to BIC.
7	Fisher Scoring	This is just showing the number of iterations the model went through before it converged to this solution (not really useful).

[4] AIC can also be used in non-nested models, but using it requires caution. The data must be exactly the same.