Lab 12: Logistic Regression

Spam Emails

Today we will be working with a corpus of emails received by a single gmail account over the first three months of 2012. Just like any other email address this account received and sent regular emails as well as receiving a large amount of spam (unsolicited bulk email). We will be using what we have learned about logistic regression models to see if we can build a model that is able to predict whether or not a message is spam based on a variety of characteristics of the email (e.g. inclusion of words like winner, inherit, or password, the number of exclamation marks used, etc.) While the spam filters used by large coorportations like Google and Microsoft are quite a bit more complex the fundamental idea is the same - binary classification based on a set of predictors.

Template for lab report

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Knit often to more easily determine the source of the error.

```
download.file("http://stat.duke.edu/~cr173/Sta102_Fa14/Lab/lab12.Rmd", destfile = "lab12.Rmd", method="wget")
```

The data

```
download.file("http://stat.duke.edu/~cr173/Sta102_Fa14/Lab/email.Rdata", destfile = "email.Rdata", method="wget")
load("email.Rdata")
```

List of variables:

1. `spam`: Indicator for whether the email was spam.
2. `to_multiple`: Indicator for whether the email was addressed to more than one recipient.
3. `from`: Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
4. `cc`: Indicator for whether anyone was CCed.
5. `sent_email`: Indicator for whether the sender had been sent an email in the last 30 days.
6. `image`: Indicates whether any images were attached.
7. `attach`: Indicates whether any files were attached.
8. `dollar`: Indicates whether a dollar sign or the word 'dollar' appeared in the email.
9. `winner`: Indicates whether "winner" appeared in the email.
10. `inherit`: Indicates whether "inherit" (or an extension, such as inheritance) appeared in the email.
11. `password`: Indicates whether "password" appeared in the email.
12. `num_char`: The number of characters in the email, in thousands.
13. `line_breaks`: The number of line breaks in the email (does not count text wrapping).
14. `format`: Indicates whether the email was written using HTML (e.g. may have included bolding or active links) or plaintext.
15. `re_subj`: Indicates whether the subject started with "Re:", "RE:", "re:", or "rE":
16. `exclaim_subj`: Indicates whether there was an exclamation point in the subject.

17. `urgent_subj`: Indicates whether the word "urgent" was in the email subject.
18. `exclaim_mess`: The number of exclamation points in the email message.
19. `number`: Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

Exercise 1 In this lab we will focus on predicting whether an email is spam or not. How many emails make up this data set? What proportion of the emails were spam?

A simple spam filter

We will start with a simple spam filter that will only use a single predictor `to_multiple` to classify a message as spam or not. To do this we will fit a logistic regression model between `spam` and `to_multiple` using the `glm` function. This is done in the same way that a simple or multiple regression model is fit in R, except we use the `glm` function instead of `lm`, and we must indicate that we wish to fit a logistic model by include the argument `family=binomial`.

```
g_simple = glm(spam ~ to_multiple, data=email, family=binomial)
summary(g_simple)
```

```
##
## Call:
## glm(formula = spam ~ to_multiple, family = binomial, data = email)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.477  -0.477  -0.477  -0.477   2.809
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.11609    0.05618 -37.665  < 2e-16 ***
## to_multipleyes -1.80918    0.29685  -6.095  1.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2437.2  on 3920  degrees of freedom
## Residual deviance: 2372.0  on 3919  degrees of freedom
## AIC: 2376
##
## Number of Fisher Scoring iterations: 6
```

Exercise 2 Based on the results of this logistic regression does the inclusion of multiple recipients make a message more or less likely to be spam? Explain your reasoning.

Exercise 3 Using these results calculate the probability that a message is spam if it has multiple recipients, what is the probability if it does not have multiple recipients.

Exercise 4 Pick one of the other 17 remaining predictors that you think is most likely to predict an email's spam status and fit a `glm` model using it. Describe how this predictor affects the probability of an email being spam.

A more complex spam filter

We will now fit the fully specified model by including all the predictors in our model. Note that we use "." as short hand to indicate that we wish to include all predictors in our model. For the time being we will ignore the warning that is produced when we fit the model.

```
g_full = glm(spam ~ ., data=email, family=binomial)
```

```
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

```
summary(g_full)
```

```
##
## Call:
## glm(formula = spam ~ ., family = binomial, data = email)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0005  -0.4368  -0.1746   0.0000   3.3902
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.998e+01  5.917e+03   0.003 0.997305
## to_multipleyes  -2.639e+00  3.314e-01  -7.962 1.69e-15 ***
## fromyes         -2.065e+01  5.917e+03  -0.003 0.997216
## ccyes           -5.978e-01  3.409e-01  -1.754 0.079474 .
## sent_emailyes   -1.710e+01  2.838e+02  -0.060 0.951954
## imageyes        -1.716e+00  8.009e-01  -2.143 0.032105 *
## attachyes        1.278e+00  2.706e-01   4.723 2.33e-06 ***
## dollaryes        1.672e-01  1.808e-01   0.925 0.355164
## winneryes        1.924e+00  3.612e-01   5.325 1.01e-07 ***
## inheityes        3.078e-02  3.399e-01   0.091 0.927852
## passwordyes     -1.788e+00  5.421e-01  -3.298 0.000973 ***
## num_char         3.033e-02  2.423e-02   1.252 0.210647
## line_breaks     -4.771e-03  1.350e-03  -3.533 0.000411 ***
## formatPlain      6.400e-01  1.505e-01   4.252 2.12e-05 ***
## re_subjyes      -1.403e+00  4.109e-01  -3.415 0.000638 ***
## exclaim_subjyes -1.409e-02  2.412e-01  -0.058 0.953433
## urgent_subjyes   3.584e+00  1.244e+00   2.881 0.003965 **
## exclaim_mess     9.642e-03  1.783e-03   5.409 6.34e-08 ***
## numbersmall     -1.233e+00  1.577e-01  -7.820 5.28e-15 ***
## numberbig       -4.623e-01  2.299e-01  -2.011 0.044299 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2437.2  on 3920  degrees of freedom
## Residual deviance: 1668.5  on 3901  degrees of freedom
## AIC: 1708.5
##
## Number of Fisher Scoring iterations: 18
```

Exercise 5 Imagine we are using this logistic regression model as a spam filter, every new message you receive is analyzed and its characteristics calculated for each of the 18 predictor variables. If the message contained an image is it more or less likely to be flagged as spam? What if it only had

an attachment? What if it had an image and an attachment?

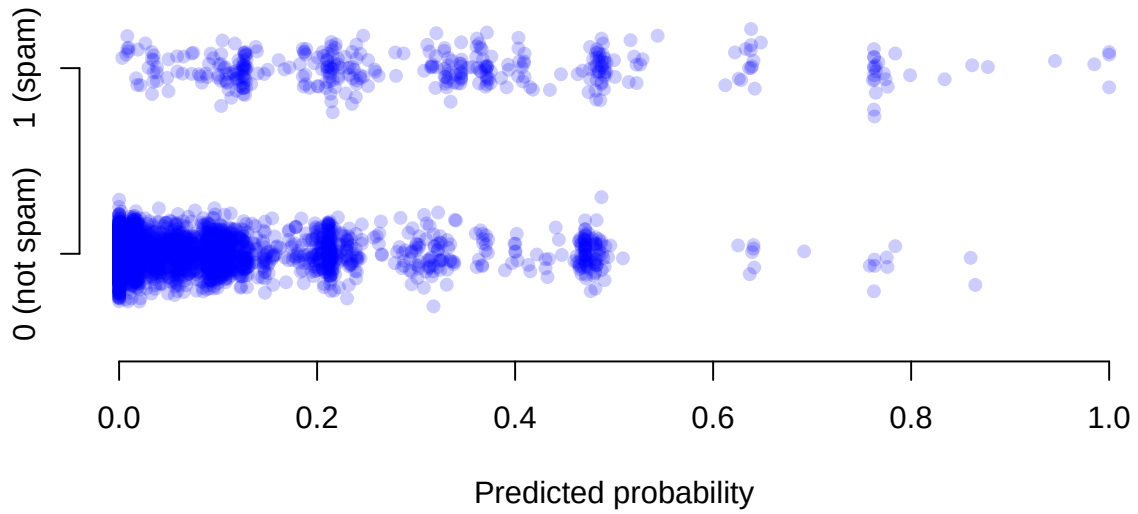
Exercise 6 Which variables appear to be meaningful for identifying spam? Describe how each of these variables affects the probability of an email being spam, numerical answers are not needed here. (For categorical predictors be sure to indicate the reference level and how this affects the interpretation)

Exercise 7 Which predictor appears to have the largest effect? Does this agree with what you have seen in the spam you receive?

Assessing our spam filter

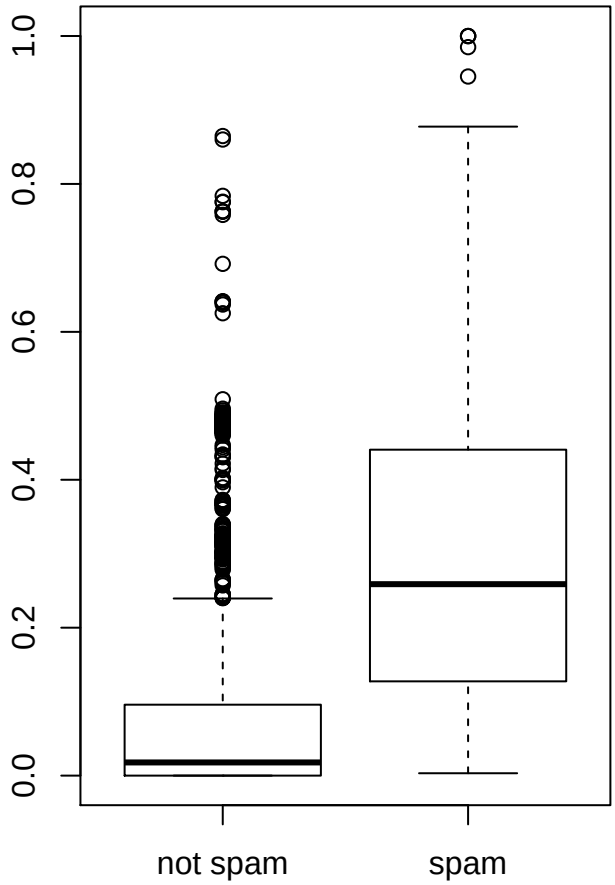
While not quite the same as the residual plots we saw in simple and multiple linear regression we can create a plot that shows how well out logistic regession model is doing by plotting our predicted probability against the response variable (spam or not spam). We jitter the y coordinates slightly so that emails with similar probabilities are not directly on top of one another in the plot.

```
set.seed(1)
jitter = rnorm(nrow(email), sd=0.08)
plot(g_full$fitted.values, email$spam+jitter,
     xlim=0:1, ylim=c(-0.5,1.5), axes=FALSE,
     xlab="Predicted probability", ylab="",
     col=adjustcolor("blue", 0.2), pch=16)
axis(1)
axis(2, at=c(0,1), labels=c("0 (not spam)", "1 (spam)"))
```



We can also see the difference in the distribution of probabilities for the two classes by plotting side by side box plots.

```
plot(factor(email$spam,labels=c("not spam","spam")),g_full$fitted.values)
```



From both plots it is clear that in general spam messages have higher probabilities, which is what we expect to see if our spam filter is working well.

Exercise 8 Fit another `glm` model with the subset of the predictors that you think will result in the best possible spam filter. Your criteria should be based on both what you know about spam and the results of fitting the full model. Describe why you chose this particular model and how its results differ from the full model.

Exercise 9 Recreate the above plots for your new model, based on these plots discuss if your model appears to be a better or worse spam filter than the full model. What other information do you think you would need to better make this decision?

Extra Credit In the dot plot above, there appears to be a banding pattern in the predicted probabilities, many points falling at the same or very nearly the same predicted probabilities. Explain why this pattern is occuring.