MATH 4044
Statistics for Data Science

Logistic regression

Course outline

Statistics for Data Science

- **Descriptive Statistics**
  - Displays & Summary Measures
- **Probability**
  - Normal Distribution
- **Statistical Inference**
  - Interval Estimation
  - One-Sample Hypothesis Tests
  - Two-Sample Hypothesis Tests
  - General Linear Models
  - Non-Parametric Tests
- **Relationships in Data**
  - Correlation & Linear Regression **Logistic Regression**
  - Chi-Square Test

Week 13

Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 8, 8.1-8.8.

Topics to be covered

- Binary logistic regression
  - Predicting binary categorical outcome variables using:
    - Categorical and continuous explanatory variables.
    - Models with interactions.
  - Model fit and diagnostics.
  - Odds ratios.

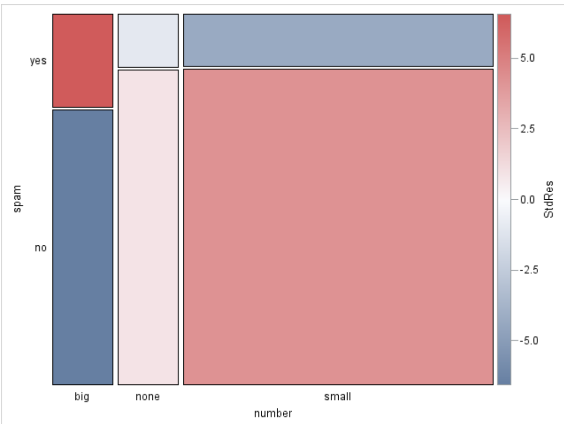
Recall: Identifying spam email

- By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy.
- We will use information about 3,921 emails collected from a single email account in early 2012 to develop a basic spam filter.
  - While our model will not be the same as those used in large-scale spam filters, it shares many of the same features.

Example: Identifying spam email

Variable	Description
spam	Specifies whether the message was spam; 1 = yes, 0 = no
num_char	The number of characters in the email, in thousands
format	Indicates if the email contained special formatting; 1 = html, 0 = text
to_multiple	Indicates if more than one person was listed in the 'To' field of the email; 1 = yes, 0 = no
cc	Indicates if someone was cc-ed on the email; 1 = yes, 0 = no
re_subj	Indicates whether 'RE:' was included at the start of the email subject; 1 = yes, 0 = no
exclaim_subj	Indicates whether an exclamation point was included in the email subject; 1 = yes, 0 = no
urgent_subj	Indicates whether the email was flagged as urgent; 1 = yes, 0 = no
winner	Indicates if the word 'winner' appeared in the email; 1 = yes, 0 = no
number	Indicates whether the email contained a number; none = no number, small = number under one million, big = large number

Example: Identifying spam email



Mosaic plot

Mosaic plots use box areas to represent the number of observations that box represents.

Each column is split proportionally according to the fraction of emails that were spam in each number category.

We can again see that the `spam` and `number` variables are associated since some columns are divided in different vertical locations than others.

Example: Identifying spam email

Table of spam by number

spam	number	Frequency	Expected	Std Residual	Cell Chi-Square	Percent	Column Percent
no	big	407	458.8	-6.5569	5.8533	10.38	74.68
	none	470	462.2	0.9851	0.1320	11.99	85.61
	small	2424	2380.0	4.2953	0.8139	61.82	85.74
	Total	3301				84.19	
yes	big	138	86.1770	6.5569	31.1640	3.52	25.32
	none	79	86.8095	-0.9851	0.7026	2.01	14.39
	small	403	447.0	-4.2953	4.3336	10.28	14.26
	Total	620				15.81	
Total	big	545				13.90	100.00
	none	549				14.00	100.00
	small	2827				72.10	100.00
	Total	3921				100.00	

Standardised residual for table cell (*i,j*):

$$\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1-p_{i\bullet})(1-p_{\bullet j})}}$$

Row / column proportion

Among emails with big numbers, the observed count of emails that are not spam is much lower than expected, resulting in a large negative standardised residual (**dark blue** area in the mosaic plot).

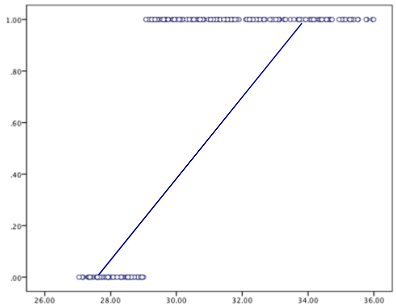
The observed count of spam emails is much higher than expected, resulting in a large positive standardised residual (**dark red** area in the mosaic plot).

Binary logistic regression

- Used to predict a binary (dichotomous) categorical response variable from one or more categorical and/or continuous explanatory variables.
- The response variable y is a dummy variable coded 0 if a condition is not present and 1 if it is.
- Instead of predicting the value of y from variable x we are interested to *predict the probability of y occurring given known values of x*.
- Thus we investigate how the *probability* that a successful outcome occurs depends upon each value of explanatory variable x.

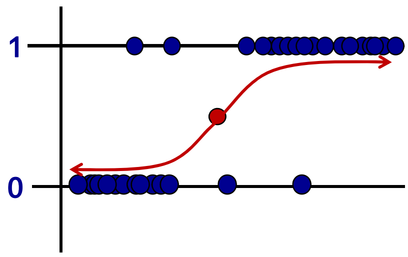
Why use binary logistic regression?

- A simple linear regression model would have a hard time fitting a straight line!
- You would typically get the correct answers in terms of the sign and significance of coefficients.



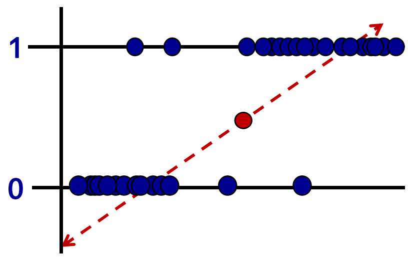
- There are three problems:
- The error terms do not have constant variance.
- The error terms are not Normally distributed.
- Probabilities are bounded between 0 and 1. If the response is coded 1 = Yes and 0 = No and your regression equation predicts 1.1 or -0.4, what does that mean?

Linear vs logistic regression



Logistic regression
Points on regression line represent predicted probability of y for each value of x

$$P(y=1) \text{ or } p$$



Linear regression
Points on regression line represent predicted transformed probability of y for each value of x

$$\ln\left(\frac{p}{1-p}\right)$$

Binary logistic regression

- The logit model can be used for binary logistic regression:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1$$

- Logit is the natural log (ln) of the odds ratio p/(1-p) with *p* the probability y takes the value 1 and *(1-p)* the probability y takes the value 0, i.e. *p* is the same as *P* (*y* = 1).
- *Why use a natural logarithm?* It transforms *y* so that we can fit an S-shaped curve with what appears to be a linear model.
- To interpret, we take the exponent *e* to "remove the ln":

Odds ratio $$\frac{p}{1-p} = e^{(b_0 + b_1 x_1)}$$

Binary logistic regression

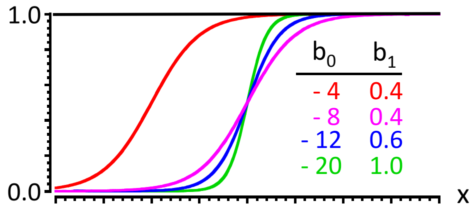
- With one explanatory variable we have *P*(*y*), the probability *y* occurs:

$$P(y=1) = \frac{e^{(b_0 + b_1 x_1)}}{1 + e^{(b_0 + b_1 x_1)}} \qquad \frac{\text{odds}}{\text{1+odds}}$$

- Where:
  - ☐ $b_1 = 0$ implies *P*(*y* = 1) is the same at each level of *x*
  - ☐ $b_1 > 0$ implies *P*(*y* = 1) increases as *x* increases
  - ☐ $b_1 < 0$ implies P(*y* = 1) decreases as *x* increases
- With several explanatory variables we still predict the probability that Y will occur:

$$P(y=1) = \frac{e^{(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m)}}{1 + e^{(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m)}}$$

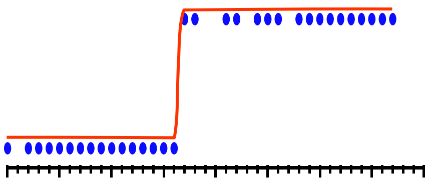
Binary logistic regression curves

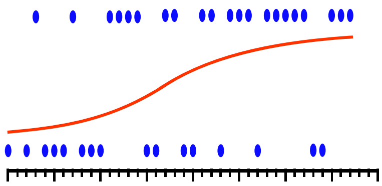


$$P(y=1) = \frac{e^{(b_0+b_1x_1)}}{1+e^{(b_0+b_1x_1)}}$$

Data with a sharp cut-off point should have a large value of b_1

Data with a lengthy transition should have a small value of b_1





Example: One categorical predictor

Is the probability an email is spam related to whether it contains any numbers?

There are no missing values

The probability being modelled is spam = 'yes'

Reference coding was used for the predictor number, with number = 'small' as reference level

Model Information

Data Set	MATH4044.EMAIL
Response Variable	spam
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	3921
Number of Observations Used	3921

Response Profile

Ordered Value	spam	Total Frequency
1	0	3301
2	1	620

Probability modeled is spam='1'.

Class Level Information

Class	Value	Design Variables	
number	big	1	0
	none	0	1
	small	0	0

Example: SAS code

We want reference coding for dummy variables with 'small' as the reference level

```
title 'logistic regression with one categorical predictor';

proc logistic data=math4044.email;

   class number (param=ref ref='small');
   model spam (event='1') = number / rsq clodds=both;

run;
quit;
```

We want to predict the probability of an email being spam

To obtain R-squared value

Produces confidence limits using the method of profile likelihood

[Requires more computation but preferred to default Wald-based intervals, particularly for sample sizes less than 50]

Example: Overall model fit

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3425.379	3390.835
SC	3431.653	3409.657
-2 Log L	3423.379	3384.835

R-Square	0.0098	Max-rescaled R-Square	0.0168

Measures used to compare competing models

Large values indicate a poorly fitting model

-2 Log L (Likelihood ratio test):
- Indicates unexplained information after model has been fitted.

AIC (Akaike information criterion) and SC (Schwartz criterion):
- Both based on log likelihood but also adjust for the number of predictors in the model.

Example: Significance of the model

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.5438	2	<.0001
Score	42.9994	2	<.0001
Wald	41.6910	2	<.0001

Different methods to assess the overall significance of the model

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
number	2	41.6910	<.0001

Assessing the significance of predictors included in the model

At 5% level of significance, we reject the null hypothesis of no relationship between the probability of an email being spam and containing numbers.

Variable `number` is a statistically significant predictor and it should be included in the logistic regression model.

Since there is only one predictor, the P-value for number is equal to the Wald value testing the global hypothesis.

Example: Parameter estimates

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7942	0.0538	1112.4278	<.0001
number	big	1	0.7127	0.1122	40.3193	<.0001
number	none	1	0.0110	0.1330	0.0068	0.9343

$$P(spam=1)=\frac{e^{(-1.7942+0.7127\times big+0.0110\times none)}}{1+e^{(-1.7942+0.7127\times big+0.0110\times none)}}$$

For an email containing big numbers:

$$P(spam=1)=\frac{e^{(-1.7942+0.7127\times 1+0.0110\times 0)}}{1+e^{(-1.7942+0.7127\times 1+0.0110\times 0)}}=0.2532$$

For an email containing no numbers:

$$P(spam=1)=\frac{e^{(-1.7942+0.7127\times 0+0.0110\times 1)}}{1+e^{(-1.7942+0.7127\times 0+0.0110\times 1)}}=0.1439$$

Example: Model testing

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	19.5	Somers' D	0.099
Percent Discordant	9.6	Gamma	0.341
Percent Tied	70.9	Tau-a	0.026
Pairs	2046620	c	0.550

- Determining concordant, discordant and tied pairs:
  - Consider all possible pairs of emails in which one is spam and the other is not.
  - For each pair, compute the probability of being spam using the model.
  - If the prediction is in the same direction as the actual pair, the pair is considered concordant. If not, the pair is considered discordant.
  - If the predicted probabilities are the same, the pair is called tied.

Example: Model testing

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	19.5	Somers' D	0.099
Percent Discordant	9.6	Gamma	0.341
Percent Tied	70.9	Tau-a	0.026
Pairs	2046620	c	0.550

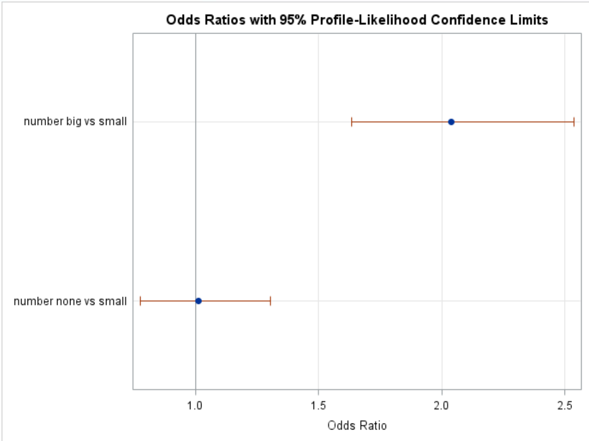
Measures of rank correlation

The higher the value, the better the predictive ability of the model

- The *c* statistic:
  - Estimates the probability that an observation with the outcome of interest will have a higher estimated probability than an observation without the outcome of interest.
  - In this example, this probability is 0.550, marginally better than chance.

Example: Odds ratios

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
number big vs small	1.0000	2.039	1.633	2.536
number none vs small	1.0000	1.011	0.775	1.305



Compared to an email with small numbers, the odds that an email with big numbers is spam are 2.039 times higher.

Emails with no numbers have the same odds of being spam as emails with small numbers.

Assumption checking

- Assumptions from Linear Regression
  - **L**inearity, **I**ndependence, **N**ormality, **E**rror variance
- Unique Problems
  - Incomplete Information: ensure all data properly collected, since combination of characteristics is important.
  - Complete Separation: when explanatory variable(s) perfectly separates the data between 0 and 1 and there is no unique model. Collecting more data can help solve this problem.
  - Overdispersion: when the observations in y have a variance larger than expected. It can occur for various reasons, such as an inadequate model specification or the observations in y are correlated with each other.

Example: One numerical predictor

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3425.379	3418.312
SC	3431.653	3430.860
-2 Log L	3423.379	3414.312

R-Square	0.0023	Max-rescaled R-Square	0.0040

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.0673	1	0.0026
Score	7.9563	1	0.0048
Wald	7.9493	1	0.0048

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5691	0.0556	795.3737	<.0001
num_char	1	-0.0103	0.00365	7.9493	0.0048

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	47.9	Somers' D	0.053
Percent Discordant	42.6	Gamma	0.058
Percent Tied	9.5	Tau-a	0.014
Pairs	2046620	c	0.526

Is the probability an email is spam related to the number of characters in the email?

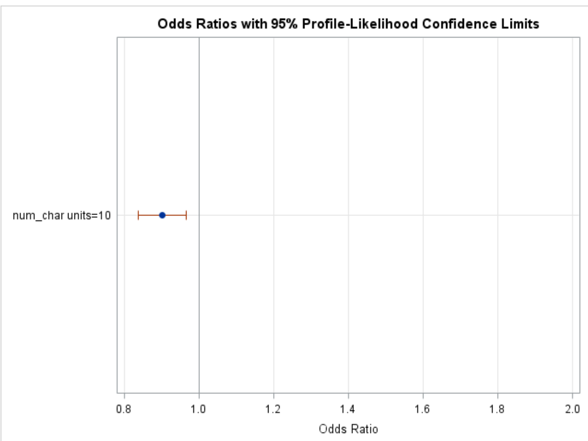
At 5% level of significance, we reject the null hypothesis of no relationship between the probability of an email being spam and the number of characters.

The number of tied pairs is much lower than for the model with number as the only predictor.

But, the value of the c statistic is only 0.526, so the predictive power of this model is still low.

Example: Odds ratio

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
num_char	10.0000	0.902	0.837	0.966



For every 10 units (thousands) increase in the number of characters, the increase in the odds ratio is 0.902 with 95% CI (0.837, 0.966).

Hence, the larger the number of characters in an email, the lower the chances the email is spam.

Example: SAS code

Because the model does not have categorical variables, this program does not include a CLASS statement.

```
title 'logistic regression with a numerical predictor only';

proc logistic data=math4044.email;
   model spam (event='1') = num_char / rsq clodds=pl;
   units num_char = 10;

run;
quit;
```

We want to estimate odds for each increase of 10 units in the number of characters

Without a UNITS statement, the odds ratio shown will be for one unit increase in the number of characters

Example: Model with interactions

Is the probability an email is spam related to its format and whether it contains numbers?

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3425.379	3013.862
SC	3431.653	3051.506
-2 Log L	3423.379	3001.862

R-Square	0.1019	Max-rescaled R-Square	0.1750

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	421.5171	5	<.0001
Score	454.2265	5	<.0001
Wald	377.4055	5	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
format	1	330.6434	<.0001
number	2	49.6048	<.0001
format*number	2	40.6427	<.0001

Model fit statistics for the model including number and format plus interactions are lower than for the model with number alone, indicating a *better fit*.

At 5% level of significance, we reject the null hypothesis of no relationship between the probability of an email being spam and the number of characters.

Effects `format`, `number` and the interaction term `format*number` are all statistically significant.

Example: Model with interactions

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			1	-0.5506	0.0763	52.1127	<.0001
format	1		1	-2.1442	0.1179	330.6434	<.0001
number	big		1	0.2968	0.2128	1.9458	0.1630
number	none		1	-1.0691	0.1627	43.1779	<.0001
format*number	1	big	1	1.0755	0.2588	17.2684	<.0001
format*number	1	none	1	1.6323	0.2964	30.3372	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	62.9	Somers' D	0.474
Percent Discordant	15.5	Gamma	0.604
Percent Tied	21.6	Tau-a	0.126
Pairs	2046620	c	0.737

This is a *hierarchical model* - main effects cannot be removed from the model if these effects are involved in an interaction that remains in the model.

The number of concordant pairs is much higher and the number of tied pairs is much lower than for the model with number as the only predictor.

The value of the *c* statistic is 0.737, so the predictive power of this model is much higher.

Example: Model with interactions

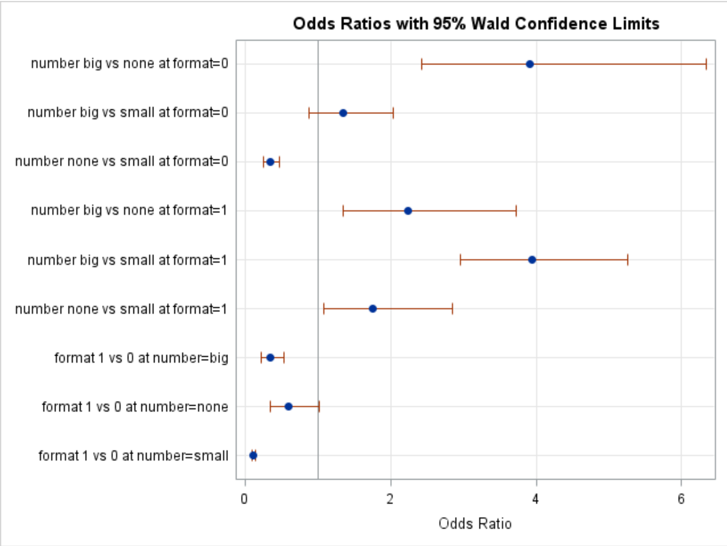
Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
number big vs none at format=0	3.919	2.424	6.338
number big vs small at format=0	1.346	0.887	2.042
number none vs small at format=0	0.343	0.250	0.472
number big vs none at format=1	2.246	1.353	3.729
number big vs small at format=1	3.945	2.955	5.265
number none vs small at format=1	1.756	1.081	2.854
format 1 vs 0 at number=big	0.343	0.219	0.539
format 1 vs 0 at number=none	0.599	0.352	1.021
format 1 vs 0 at number=small	0.117	0.093	0.148

Plain text emails with numbers have much higher odds of being spam that plain text emails without numbers at all.

For emails in html format, emails with big numbers have much higher odds of being spam.

Compared to plain text emails, those in html format are less likely to be spam, with the lowest odds of being spam for html emails with small numbers.

Example: Model with interactions



Example: SAS code

```
title 'logistic regression with interactions';

proc logistic data=math4044.email;
   class
   format (ref='0')
   number (ref='small')/ param=ref;
   model spam (event='1') = format | number / rsq clodds=pl;

   oddsratio number;
   oddsratio format;

run;
quit;
```

Variables `format` and `number` plus interactions

When there are interactions in the model, proc logistic does not automatically compute odds ratios. ODDSRATIO statement is used to obtain estimates of odds ratios.

Example: Model based on backwards selection

Which characteristics together help predict whether an email is or is not spam?

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3425.379	3102.469
SC	3431.653	3140.114
-2 Log L	3423.379	3090.469

R-Square	0.0814	Max-rescaled R-Square	0.1398

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	332.9095	5	<.0001
Score	351.9888	5	<.0001
Wald	311.2253	5	<.0001

Final model with `format`, `re_subj`, `exclaim_subj`, `cc`, and `num_char` as predictors of probability an email is spam

At 5% level of significance, we reject the null hypothesis of no relationship between the probability of an email being spam and the five predictors.

Example: Model based on backwards selection

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	to_multiple	1	7	0.0012	0.9724
2	urgent_subj	1	6	0.0108	0.9171
3	winner	1	5	2.3182	0.1279

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
format	1	267.3672	<.0001
re_subj	1	28.9020	<.0001
exclaim_subj	1	7.7174	0.0055
cc	1	10.1785	0.0014
num_char	1	8.4751	0.0036

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.7876	0.0701	126.3776	<.0001
format	1	1	-1.6415	0.1004	267.3672	<.0001
re_subj	1	1	-0.7817	0.1454	28.9020	<.0001
exclaim_subj	1	1	0.4407	0.1586	7.7174	0.0055
cc	1	1	0.4285	0.1343	10.1785	0.0014
num_char		1	0.00901	0.00309	8.4751	0.0036

All remaining predictors are statistically significant at 5% level.

Example: Model based on backwards selection

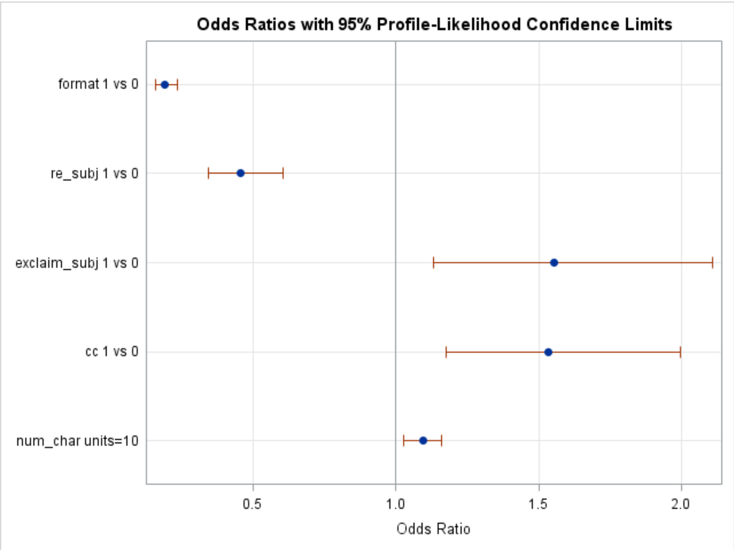
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.8	Somers' D	0.436
Percent Discordant	27.3	Gamma	0.444
Percent Tied	1.9	Tau-a	0.116
Pairs	2046620	c	0.718

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
format 1 vs 0	1.0000	0.194	0.159	0.236
re_subj 1 vs 0	1.0000	0.458	0.343	0.606
exclaim_subj 1 vs 0	1.0000	1.554	1.131	2.109
cc 1 vs 0	1.0000	1.535	1.178	1.995
num_char	10.0000	1.094	1.027	1.161

Html format and 'RE:' on the subject line correspond to much lower odds that an email is spam.

In contrast, exclamation marks on the subject line and cc increase the odds that an email is spam.

Example: Model based on backwards selection



Example: SAS code

```
proc logistic data=math4044.email;

   class to_multiple (ref='0') winner (ref='no')
         format (ref='0') re_subj (ref='0')
         exclaim_subj (ref='0') urgent_subj (ref='0')
         cc (ref='0') number (ref='small') / param=ref;

   model spam (event='1') = to_multiple winner format
         re_subj exclaim_subj urgent_subj cc num_char /
         selection=backward rsq clodds=pl;

units num_char=10;

run;
quit;
```