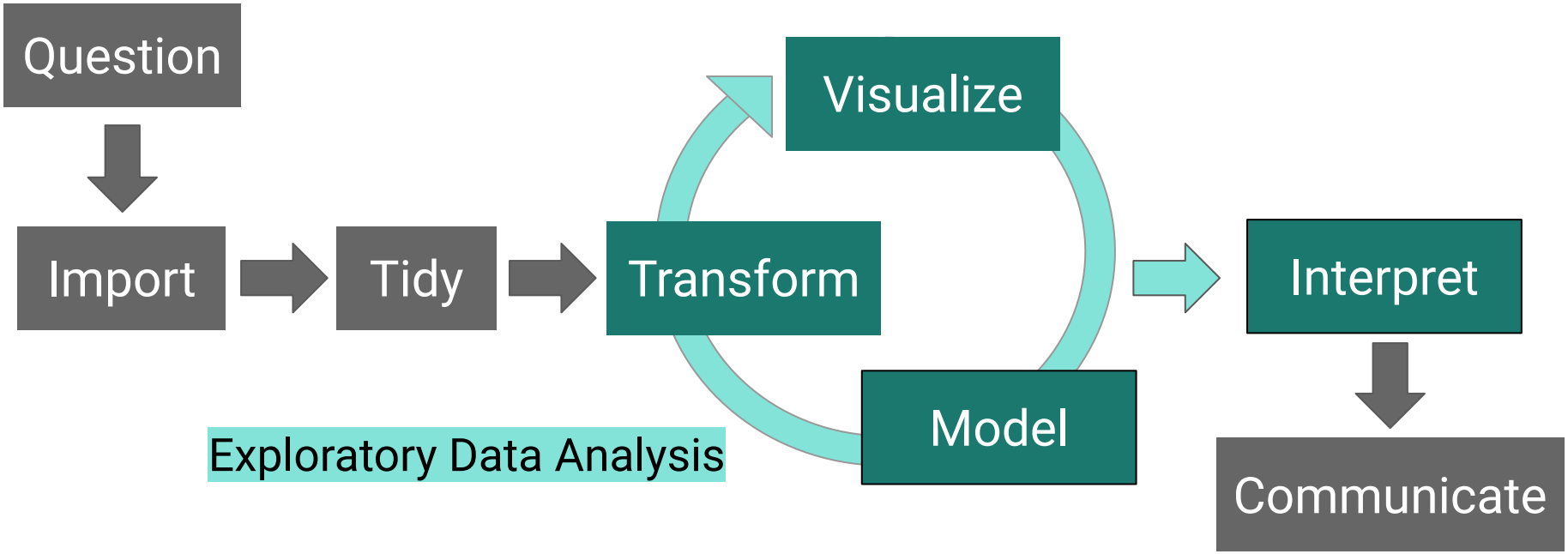
Statistical Modeling III

Lecture 13

Motivation



Modified from https://r4ds.had.co.nz/explore-intro.html#explore-intro

Regression models so far...

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Multiple regression

- Relationship between numerical response and multiple numerical and/or categorical predictors

How about when the response is a categorical variable?

Outcomes with 2 variables: Survived/Died
Success/Failure
Win/Lose

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

What are the odds of rolling a 2? **1 in 5**

What are the odds of winning the lottery? **200,000,000 to 1**

North Carolina births (ncbirths)

- In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children.
- Definition of lowbirthweight (WHO) < 2500g (5.5 lbs)

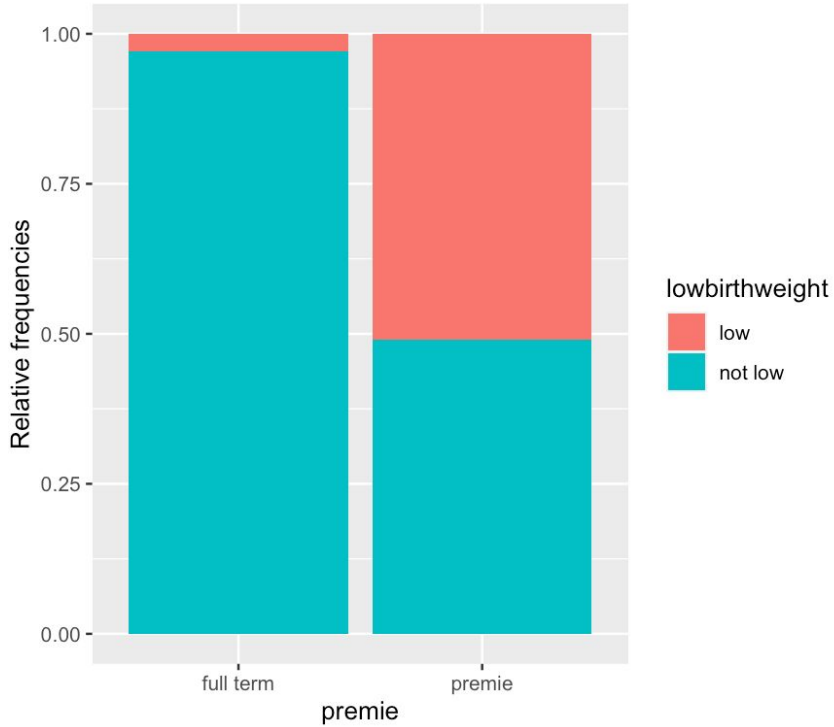
North Carolina births (ncbirths)

glimpse(ncbirths)

```
Rows: 1,000
Columns: 13
$ fage           <int> NA, NA, 19, 21, NA, NA, 18, 17, NA, 20, 30, NA, …
$ mage           <int> 13, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, …
$ mature         <fct> younger mom, younger mom, younger mom, younger m…
$ weeks          <int> 39, 42, 37, 41, 39, 38, 37, 35, 38, 37, 45, 42, …
$ premie         <fct> full term, full term, full term, full term, full…
$ visits         <int> 10, 15, 11, 6, 9, 19, 12, 5, 9, 13, 9, 8, 4, 12,…
$ marital        <fct> not married, not married, not married, not marri…
$ gained         <int> 38, 20, 38, 34, 27, 22, 76, 15, NA, 52, 28, 34, …
$ weight         <dbl> 7.63, 7.88, 6.63, 8.00, 6.38, 5.38, 8.44, 4.69, …
$ lowbirthweight <fct> not low, not low, not low, not low, not low, low…
$ gender         <fct> male, male, female, male, female, male, male, ma…
$ habit          <fct> nonsmoker, nonsmoker, nonsmoker, nonsmoker, nons…
$ whitemom       <fct> not white, not white, white, white, not white, n…
```

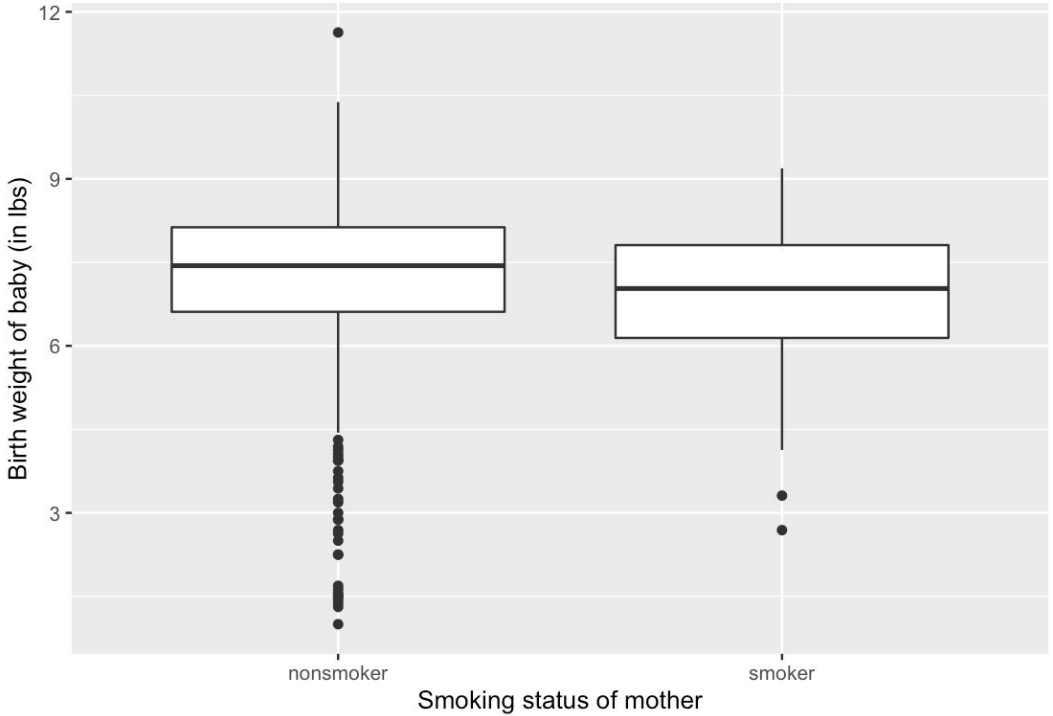
North Carolina births (ncbirths)

Lowbirthweight vs Prematurity



North Carolina births (ncbirths)

Lowbirthweight vs Habit



North Carolina births (ncbirths)

- It seems that both prematurity and smoking habits of mothers have an effect on birth weight of babies, how do we come up with a model that will let us explore this relationship?
- One way to think about the problem - we can treat Normal Birthweight and Low Birthweight as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Generalized linear models

- It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs).
- All generalized linear models have the following three characteristics:
  a. A probability distribution describing the outcome variable
  b. A linear model: $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$.
  c. A link function that relates the linear model to the parameter of the outcome distribution: $g(p) = \eta$ or $p = g^{-1}(\eta)$.

Generalized linear models

- Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.
- We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.
- To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

$$logit(p) = log(\frac{p}{1-p})$$

Generalized linear models

- Some examples of GLMs

Model	Distribution	Link Function
Logistic Regression	Binomial	Logit
Loglinear	Poisson	Log
Poisson Regression	Poisson	Log
Multinomial Response	Multinomial	Generalized Logit

Logistic regression model

The three GLM criteria:

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

To solve for p:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Multiple logistic regression model on ncbirths data

- Using tidymodels
  - Step 1: Specify the model
  - Step 2: Run the model
  - Step 3: Analyze the results

Multiple logistic regression model on ncbirths data

- Using tidymodels
  - Step 1: Specify the model
  - **logistic_reg( )** function is equivalent to **glm( )**
  - **set_engine( )** function set to **glm**

```
log_model <- logistic_reg(mode = "classification") %>%
                    set_engine("glm")
```

Multiple logistic regression model on ncbirths data

- Using tidymodels
  - Step 2: Run the model using **fit( )** function

```
# Model formula
formula <- lowbirthweight ~ habit + premie + mage + visits +
gained + gender + marital + whitemom

# Run the model
log_fit <- log_model %>%
                    fit(formula, data = ncbirths)
```

Multiple logistic regression model on ncbirths data

- Using tidymodels
  - Step 3: Access the results using the **pluck( )** and **summary( )**

```
log_model_res <- log_fit %>%
                 pluck("fit") %>%
                 summary()
log_model_res
```

Multiple logistic regression model on ncbirths data

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.42004	0.84337	-2.869	0.00411	**
habitsmoker	0.57248	0.43648	1.312	0.18966	
premiepremie	3.45586	0.31153	11.093	< 2e-16	***
mage	0.02540	0.02581	0.984	0.32520	
visits	-0.04718	0.03772	-1.251	0.21099	
gained	-0.01287	0.01155	-1.114	0.26515	
gendermale	-0.29952	0.30715	-0.975	0.32948	
maritalmarried	-0.60418	0.35356	-1.709	0.08748	.
whitemomwhite	-0.49309	0.34863	-1.414	0.15725	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple logistic regression model on ncbirths data

General model:

$$\log(p/1-p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

log(p/1-p) = -2.42 + (0.57)habit + (3.46)premie + … + (-0.49)whitemom

Nonsmoker model: (nonsmoker = 0)

log(p/1-p) = 2.42 + (0.57)(0) + (3.46)premie + … + (-0.49)whitemom
= 2.42 + (3.46)premie + … + (-0.49)whitemom

Smoker model: (smoker = 1)

log(p/1-p) = 2.42 + (0.57)(1) + (3.46)premie + … + (-0.49)whitemom
= 2.42 + 0.57 + (3.46)premie + … + (-0.49)whitemom
= 2.99 + (3.46)premie + … + (-0.49)whitemom

Confidence interval for the **premie** slope coefficient

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.42004	0.84337	-2.869	0.00411	**
habitsmoker	0.57248	0.43648	1.312	0.18966	
premiepremie	3.45586	0.31153	11.093	< 2e-16	***

...

- The interpretation for a slope is the change in log odds ratio per unit change in the predictor.
- Log odds ratio:

slope = 3.45586
CI = PE ± CV x SE
= 3.45586 ± 1.96 x 0.31153 = (2.845261, 4.066459)

- Odds ratio:

exp(slope) = exp(3.45586) = 31.68553
exp(CI) = (exp(2.845261), exp(4.066459))
= (17.20605, 58.34998)

Confidence interval for the **premie** slope coefficient

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.42004	0.84337	-2.869	0.00411	**
habitsmoker	0.57248	0.43648	1.312	0.18966	
premiepremie	3.45586	0.31153	11.093	< 2e-16	***

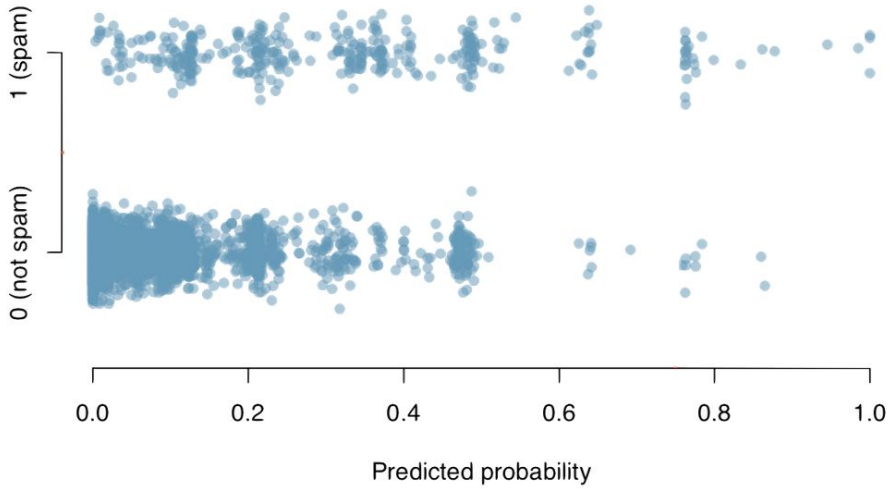
...

	Odds ratio	95% CI
premie	31.68553	(17.20605, 58.34998)

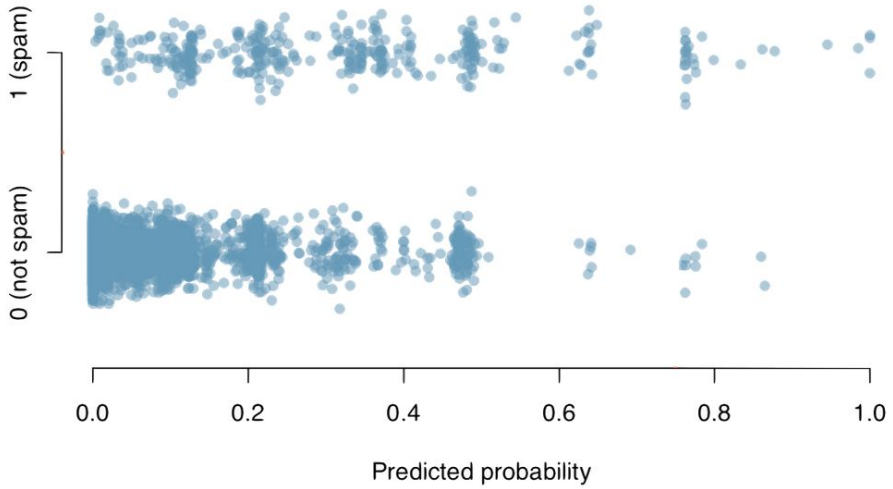
Interpretation of the slope:

- Keeping all other variables constant then, the odds ratio of getting lowbirthweight of babies from premature vs full term birth is 31.69.
- Babies from mothers with premature pregnancy are more likely to have lowbirthweights (OR 31.69, 95% CI: 17.20605, 58.34998), after controlling for all other variables.

Email spam data

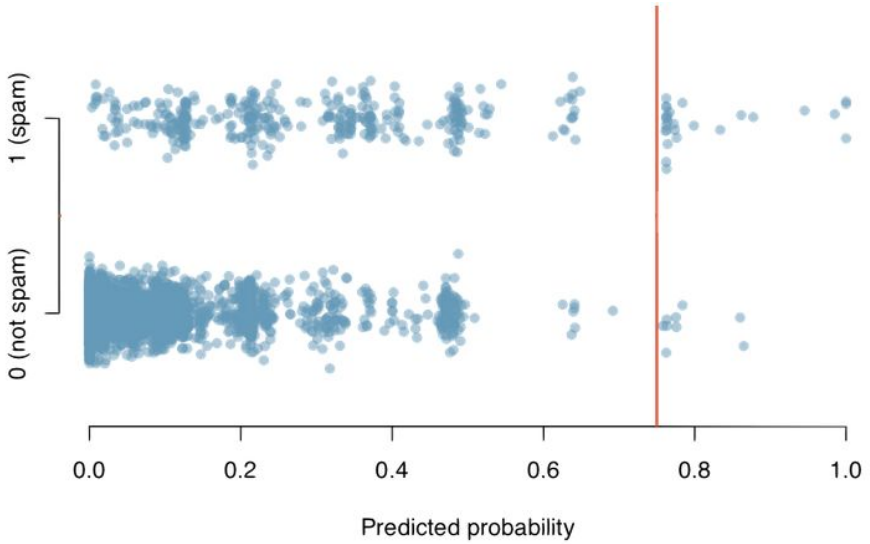


Email spam data



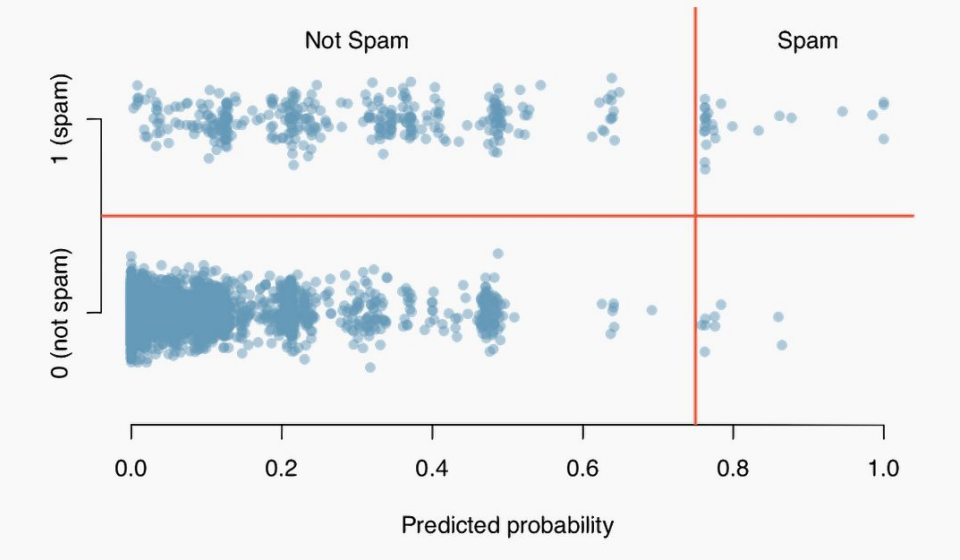
Lets see what happens if we pick our threshold to be 0.75.

Email spam data



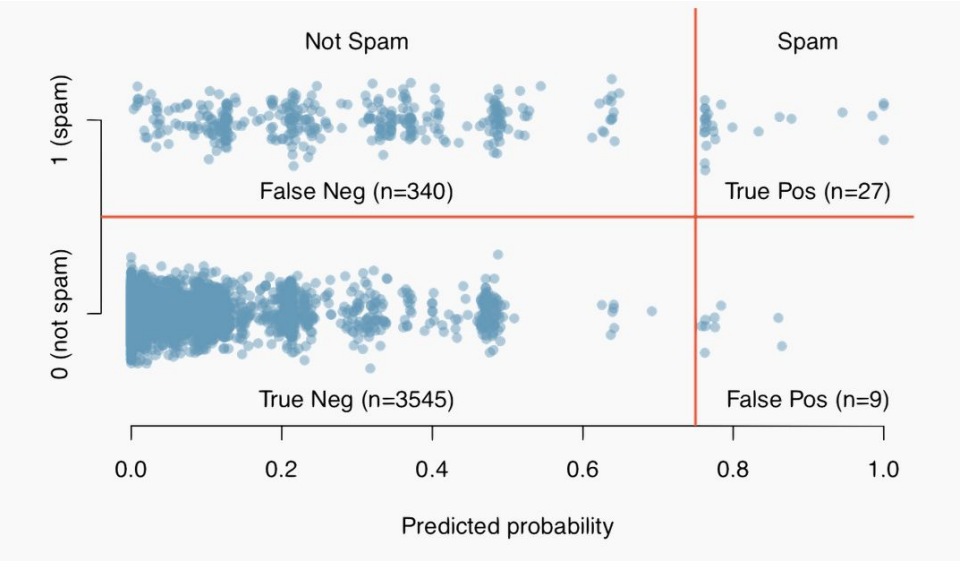
Lets see what happens if we pick our threshold to be 0.75.

Email spam data



Lets see what happens if we pick our threshold to be 0.75.

Email spam data



Lets see what happens if we pick our threshold to be 0.75.

Email spam data

For our data set picking a threshold of 0.75 gives us the following results:

FN = 340 TP = 27

TN = 3545 FP = 9

Email spam data

For our data set picking a threshold of 0.75 gives us the following results:

FN = 340 TP = 27

TN = 3545 FP = 9

What are the sensitivity and specificity for this particular decision rule?

Email spam data

For our data set picking a threshold of 0.75 gives us the following results:

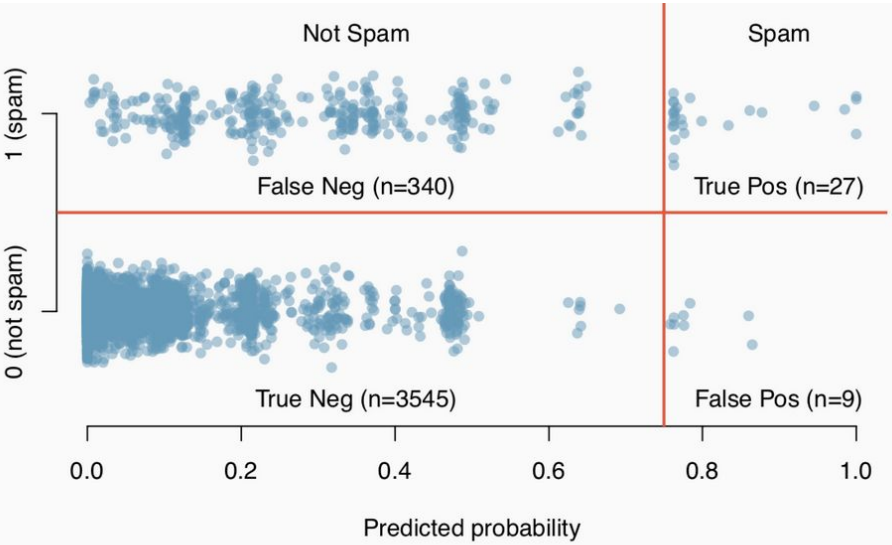
FN = 340 TP = 27

TN = 3545 FP = 9

What are the sensitivity and specificity for this particular decision rule?

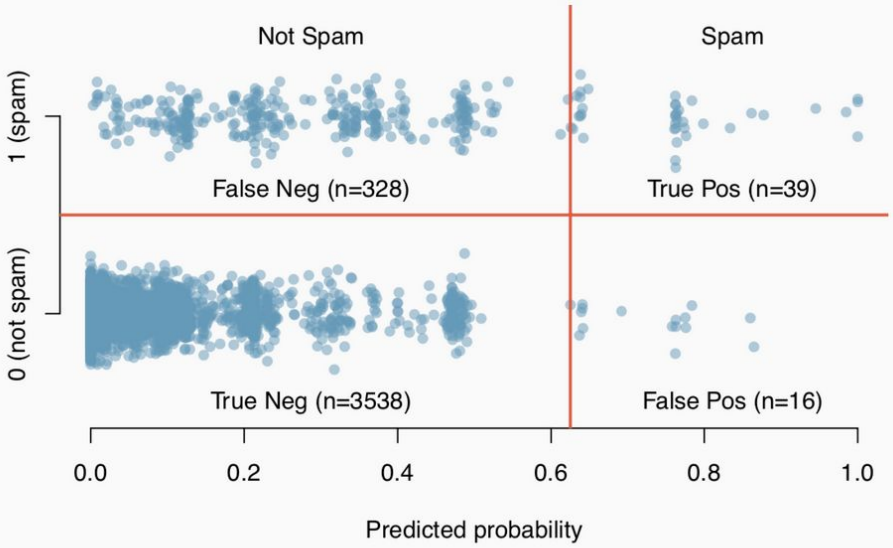
Sensitivity = TP/(TP + FN) = 27/(27 + 340) = 0.073
Specificity = TN/(FP + TN) = 3545/(9 + 3545) = 0.997

Other threshold values



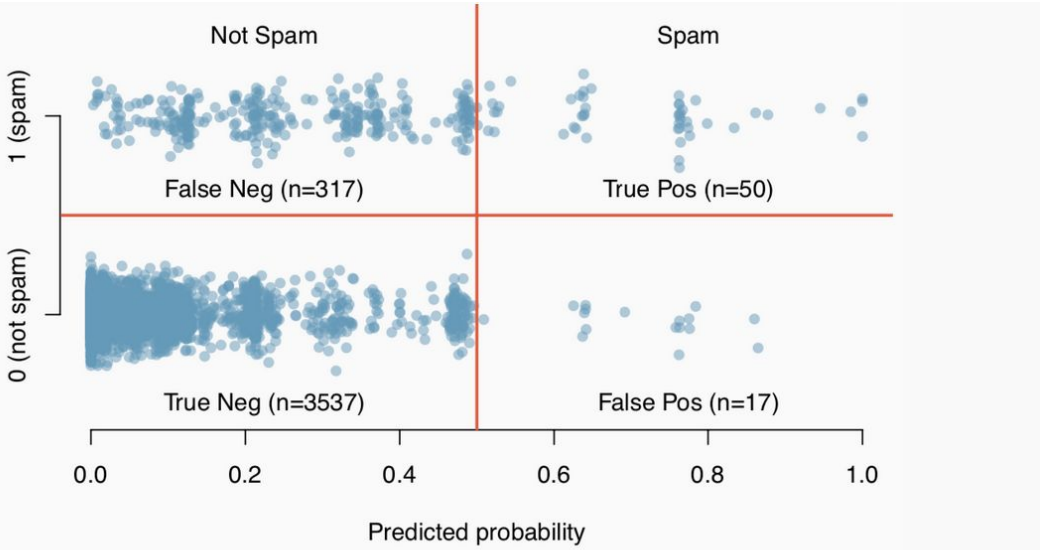
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Other threshold values



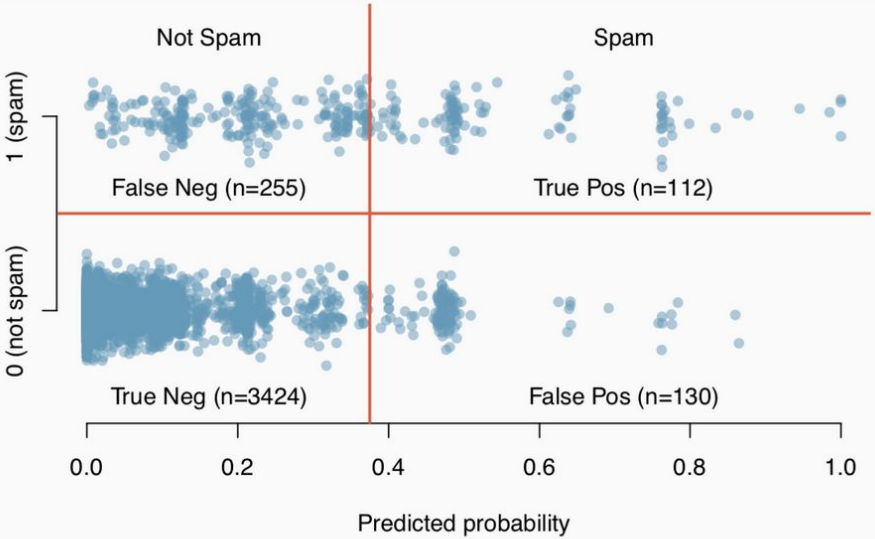
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Other threshold values



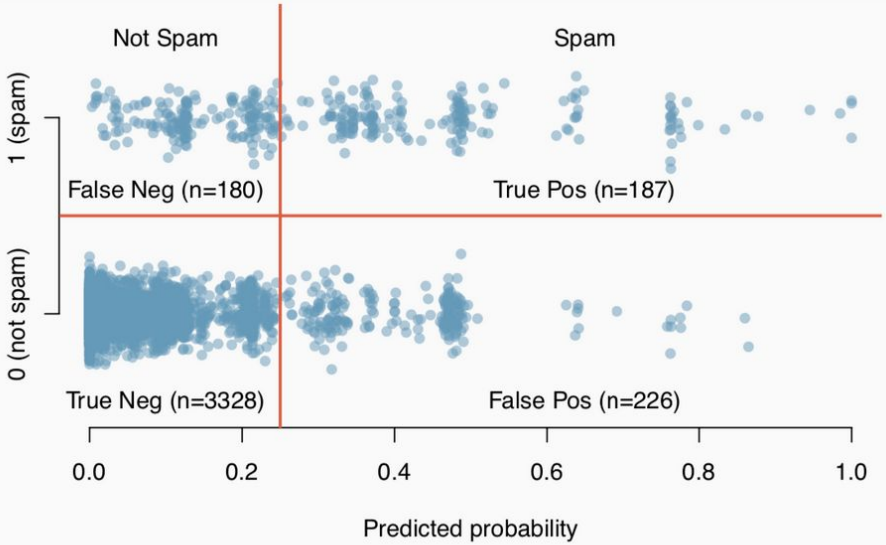
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Other threshold values



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	
Specificity	0.997	0.995	0.995	0.963	

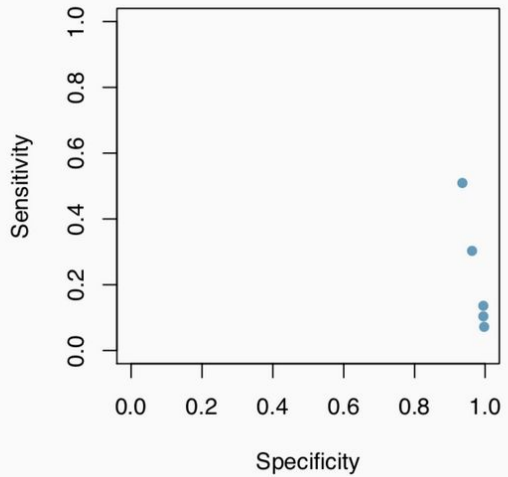
Other threshold values



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

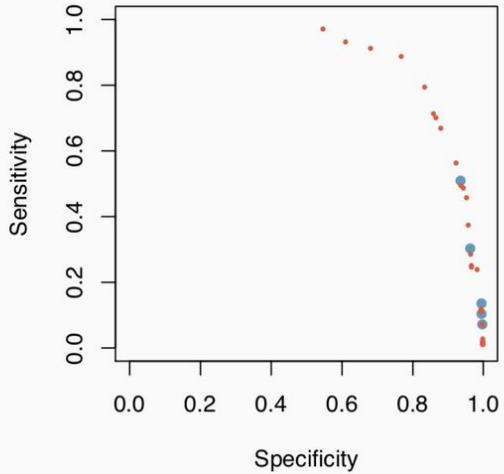
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

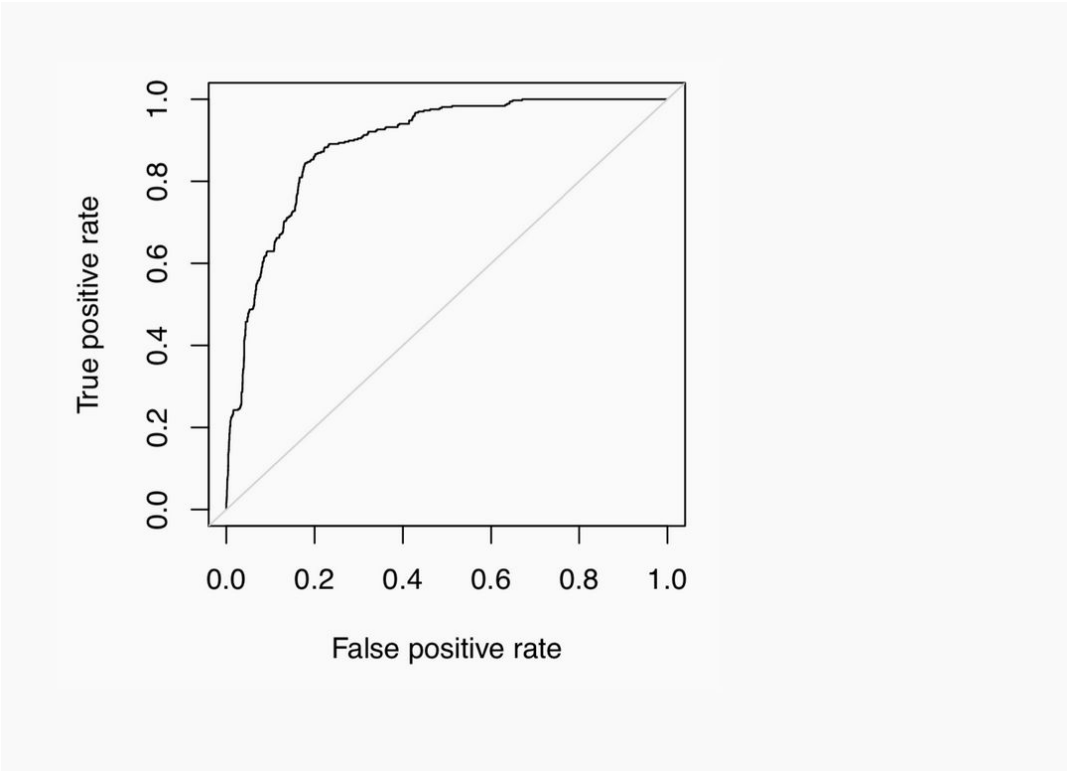


Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



Relationship between Sensitivity and Specificity

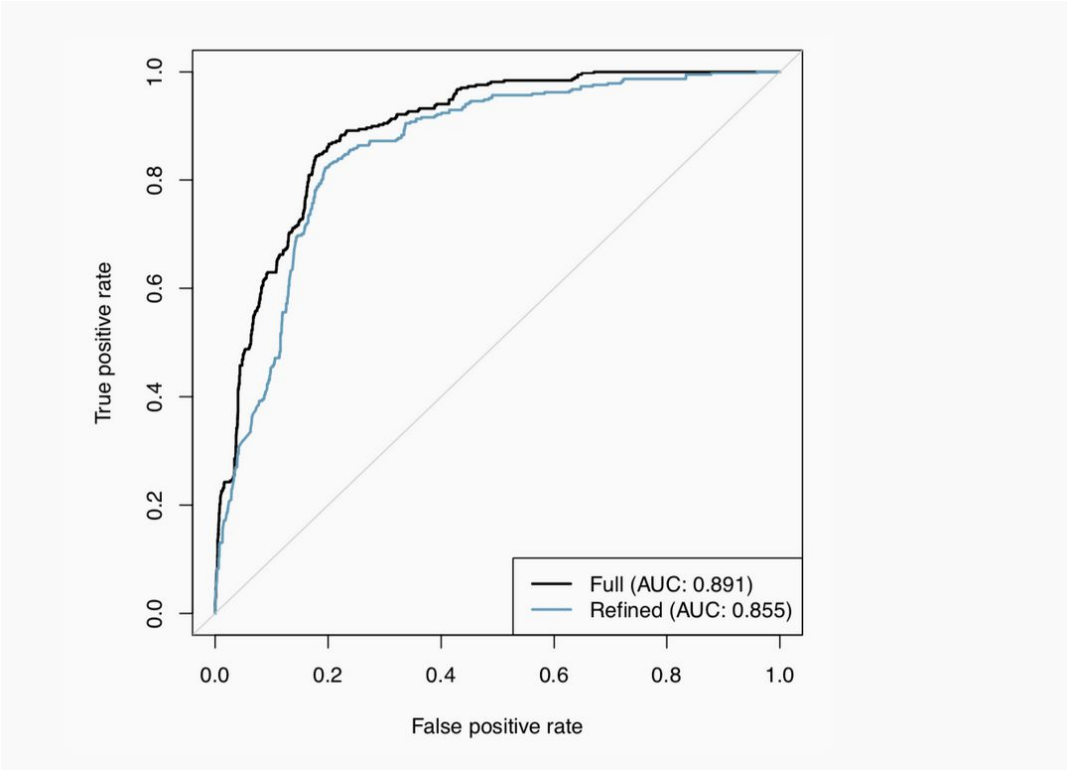


Relationship between Sensitivity and Specificity

Why do we care about ROC curves?

- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Comparing models



Takeaway message

- Use multiple logistic analysis to assess the relationship between a categorical response or outcome variable and several explanatory variables simultaneously
- ROC curves are useful for comparing several models