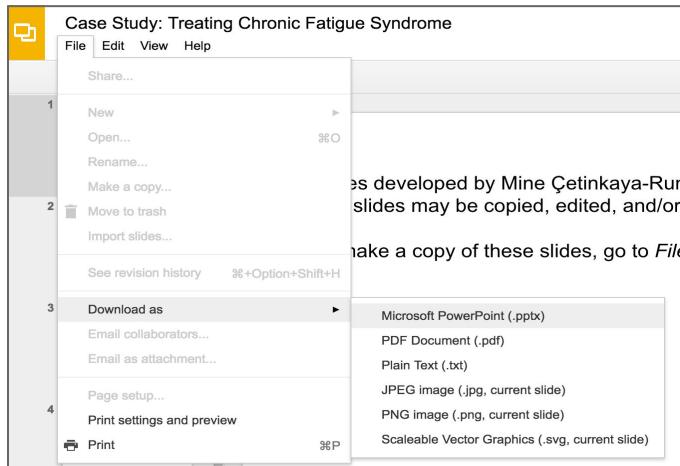


Slides developed by Mine Çetinkaya-Rundel of OpenIntro

Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.

The slides may be copied, edited, and/or shared via the [CC BY-SA license](#)

To make a copy of these slides, go to *File > Download as > [option]*, as shown below. Or if you are logged into a Google account, you can choose *Make a copy...* to create your own version in Google Drive.



Logistic regression

Regression so far ...

At this point we have covered:

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Regression so far ...

At this point we have covered:

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

Regression so far ...

At this point we have covered:

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

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Odds

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

For some event E,

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

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed)

Example - Donner Party - Data

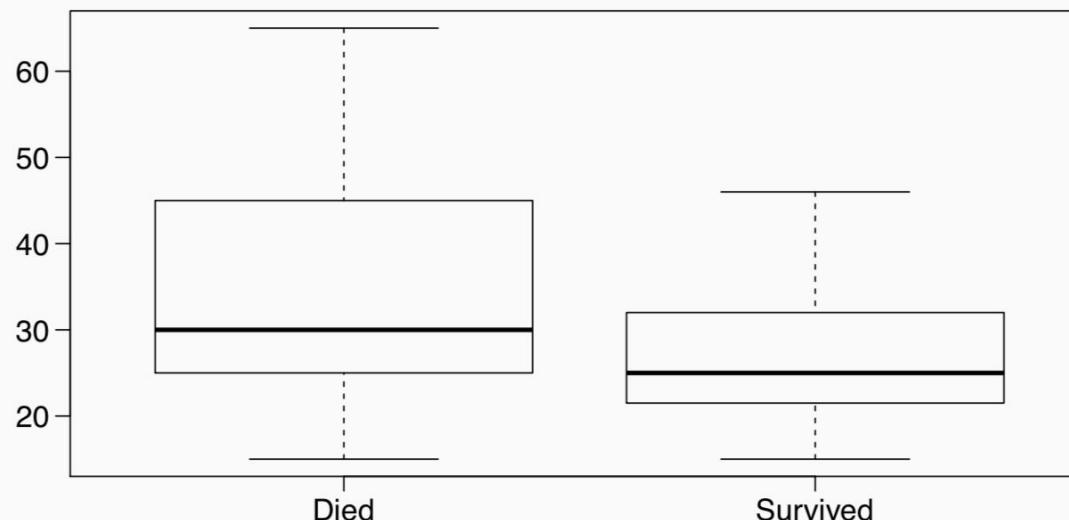
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
:	:	:	:
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs Gender

	Male	Female
Died	20	5
Survived	10	10

Status vs Age



Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

One way to think about the problem - we can treat Survived and Died 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). Logistic regression is just one example of this type of model.

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). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model: $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$.
3. A link function that relates the linear model to the parameter of the outcome distribution: $g(p) = \eta$ or $p = g^{-1}(\eta)$.

Logistic Regression

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.

Logistic Regression

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 function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞ .

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, 68 more on this later.

The logistic regression model

The three GLM criteria give us:

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

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

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

From which we arrive at,

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

Example - Donner Party - Model

In **R** we fit a GLM in the same was as a linear model except using **glm** instead of **lm** and we must also specify the type of GLM to fit using the **family** argument.

```
summary(glm(Status ~ Age, data=donner, family=binomial))
## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.81852   0.99937   1.820   0.0688 .
## Age        -0.06647   0.03222  -2.063   0.0391 *
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 56.291 on 43 degrees of freedom
## AIC: 60.291
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

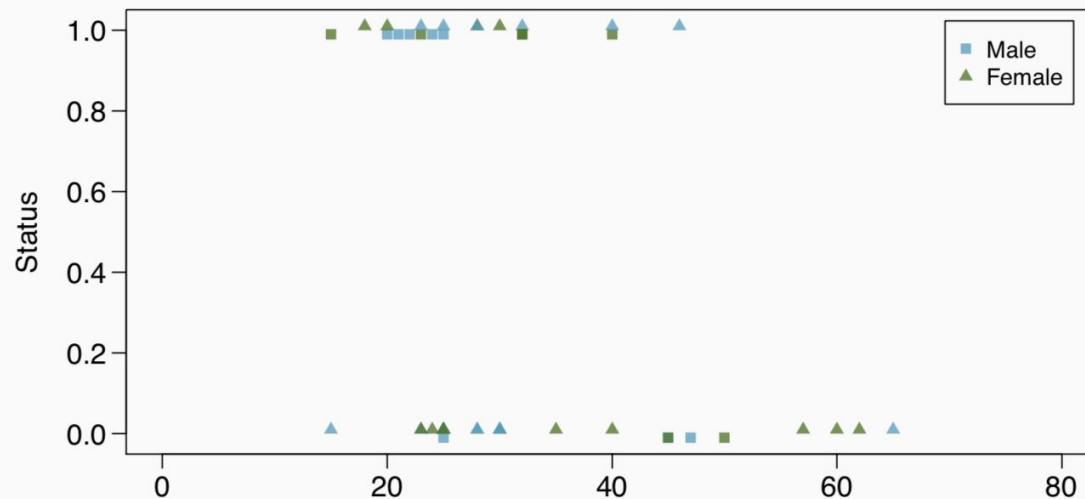
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

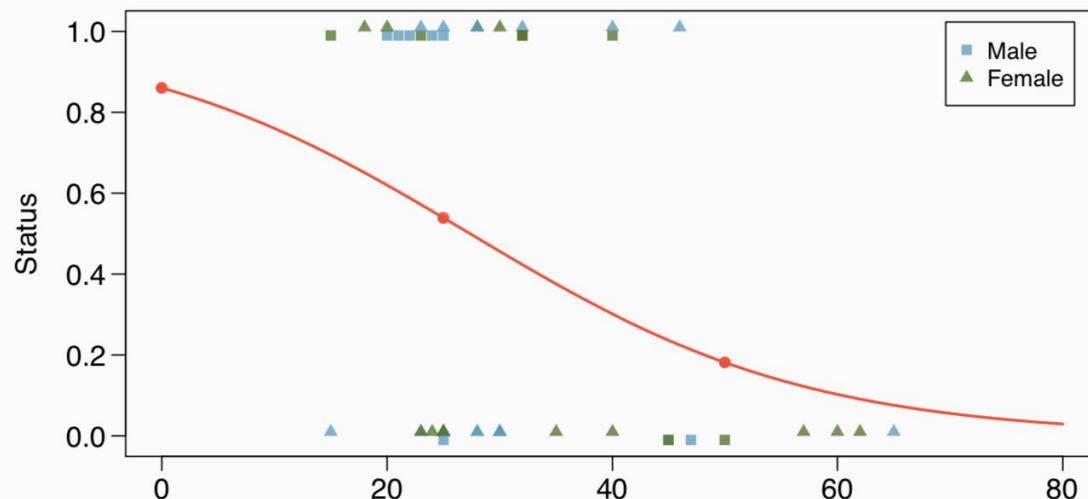
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.8185 - 0.0665(x + 1) \\ &= 1.8185 - 0.0665x - 0.0665\end{aligned}$$

$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1} \middle| \frac{p_2}{1-p_2}\right) = -0.0665$$

$$\frac{p_1}{1-p_1} \middle| \frac{p_2}{1-p_2} = \exp(-0.0665) = 0.94$$

Example - Donner Party - Age and Gender

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))
## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age         -0.07820   0.03728  -2.097   0.0359 *
## SexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## 
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

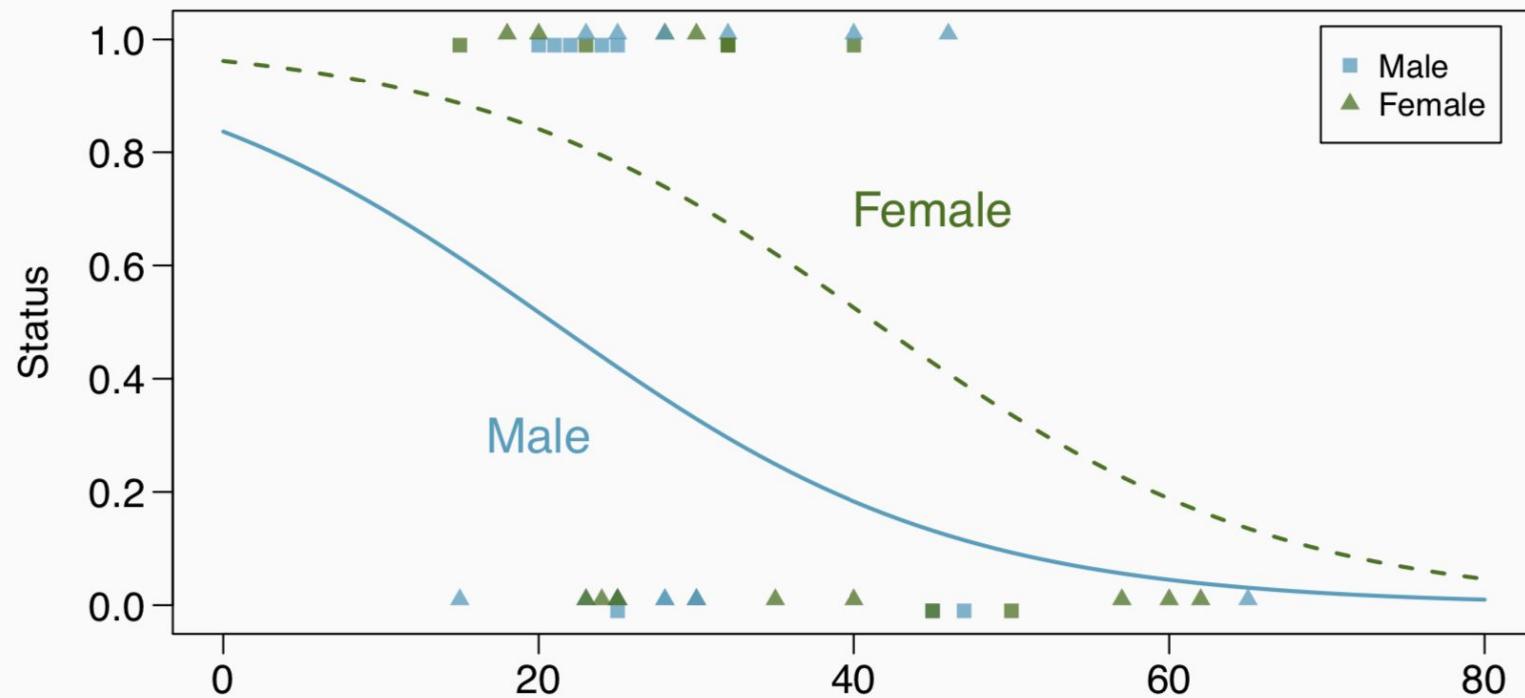
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 0 \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 1 \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age         -0.07820   0.03728  -2.097   0.0359 *
## SexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## 
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age         -0.07820   0.03728  -2.097   0.0359 *
## SexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Note: The model output does not include any F-statistic, as a general rule there are not single model hypothesis tests for GLM models.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z-test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta}_{age} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\begin{aligned} p\text{-value} &= P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10) \\ &= 2 \times 0.0178 = 0.0359 \end{aligned}$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.8596, 0.9949)$$

Example - Birdkeeping and Lung Cancer

A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

From Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed)

Example - Birdkeeping and Lung Cancer - Data

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
:	:	:	:	:	:	:	:
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

LC Whether subject has lung cancer

FM Sex of subject

SS Socioeconomic status

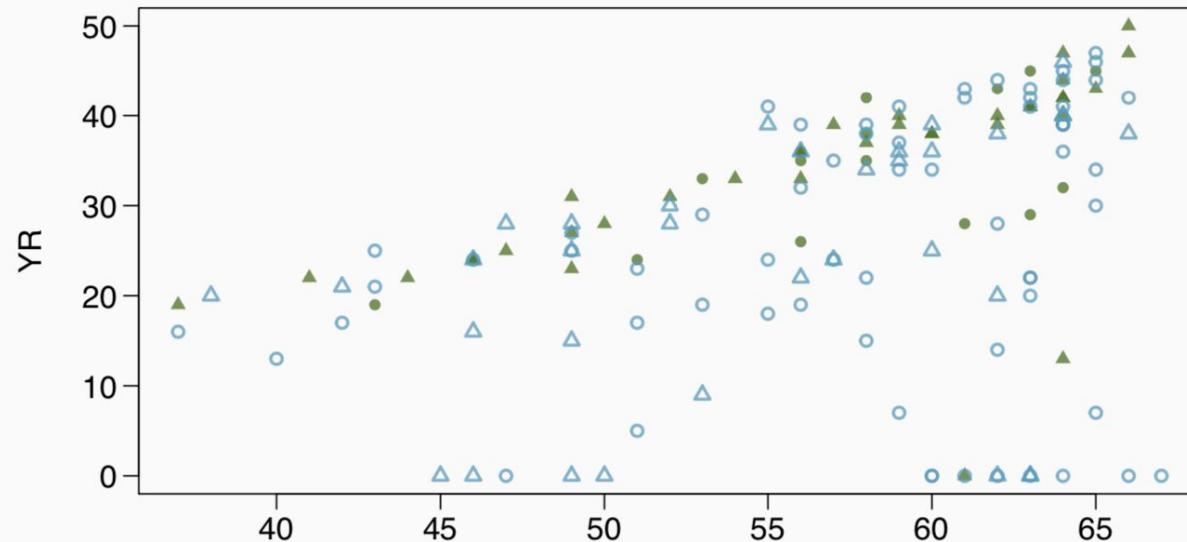
BK Indicator for birdkeeping

AG Age of subject (years)

YR Years of smoking prior to diagnosis or examination

CD Average rate of smoking (cigarettes per day)

Example - Birdkeeping and Lung Cancer - EDA



	Bird	No Bird
Lung Cancer	▲	●
No Lung Cancer	△	○

Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))
## Call:
## glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = binomial,
##      data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736  1.80425   -1.074 0.282924
## FMFemale     0.56127  0.53116    1.057 0.290653
## SSHigh       0.10545  0.46885    0.225 0.822050
## BKBird       1.36259  0.41128    3.313 0.000923 ***
## AG           -0.03976  0.03548   -1.120 0.262503
## YR            0.07287  0.02649    2.751 0.005940 **
## CD            0.02602  0.02552    1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
```

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.3626) = 3.91$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0729) = 1.08$.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are not 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between relative risk and an odds ratio.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

Back to the birds

What is probability of lung cancer in a bird keeper if we knew that
 $P(\text{lung cancer}|\text{no birds}) = 0.05$?

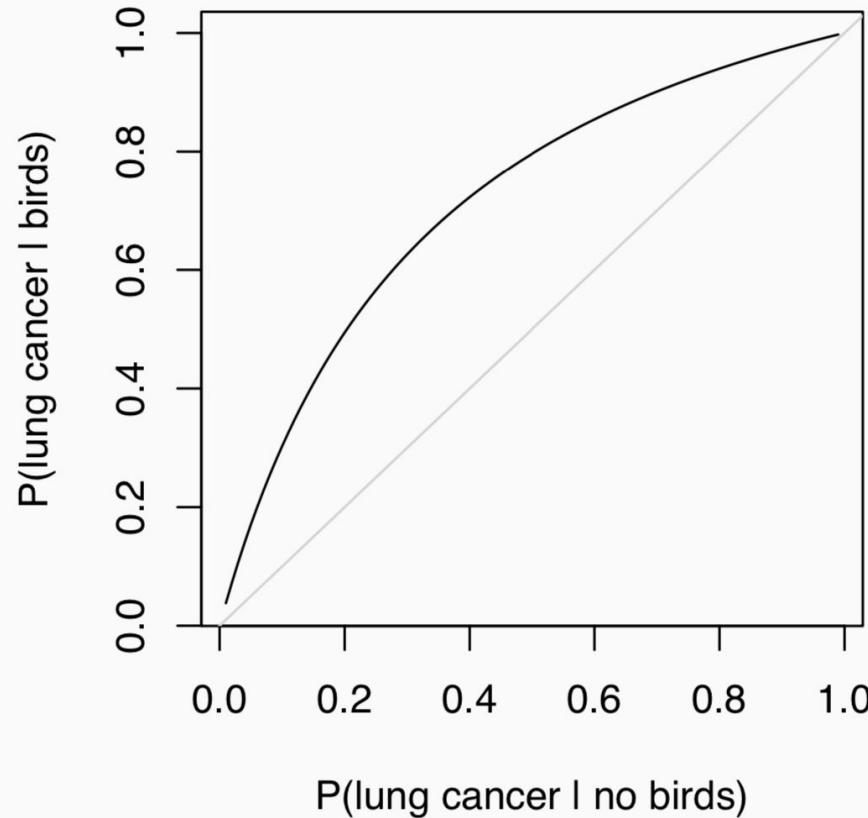
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 3.91$$

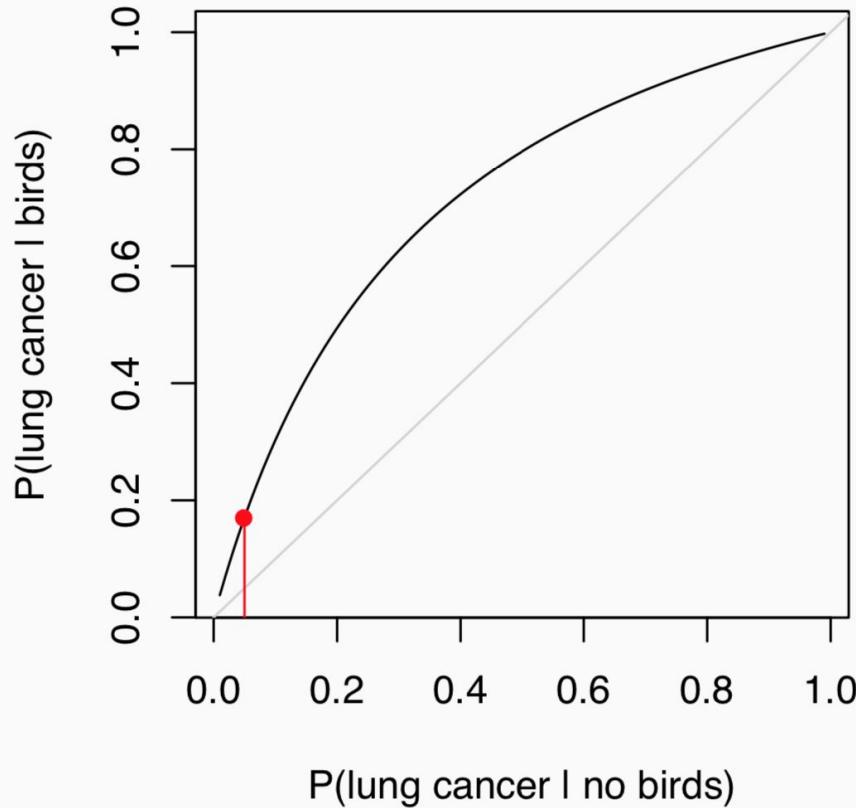
$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.171/0.05 = 3.41$$

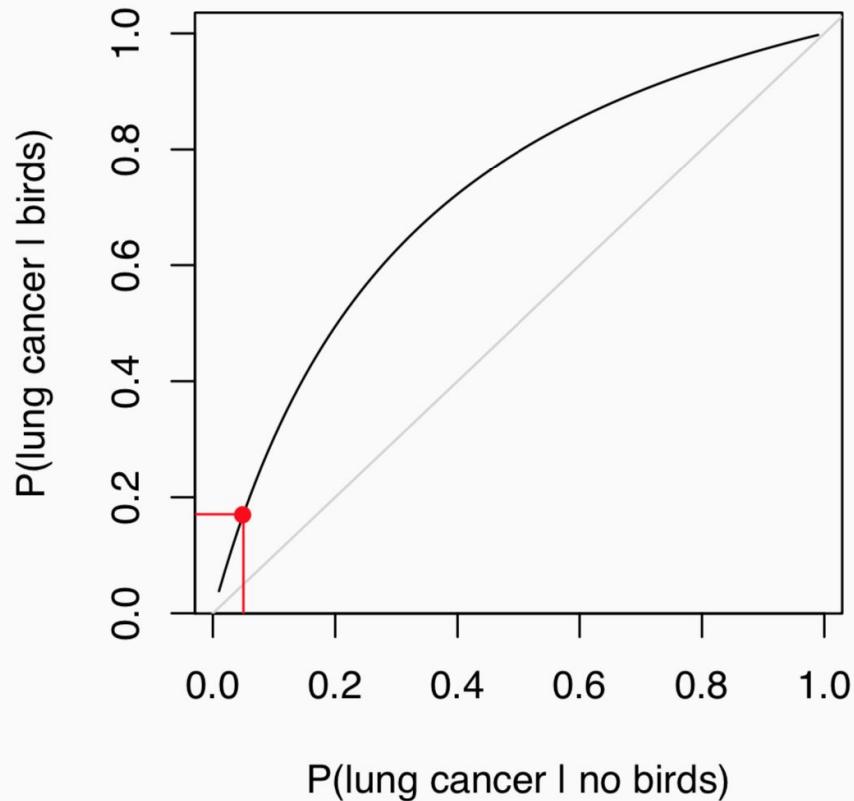
Bird OR Curve



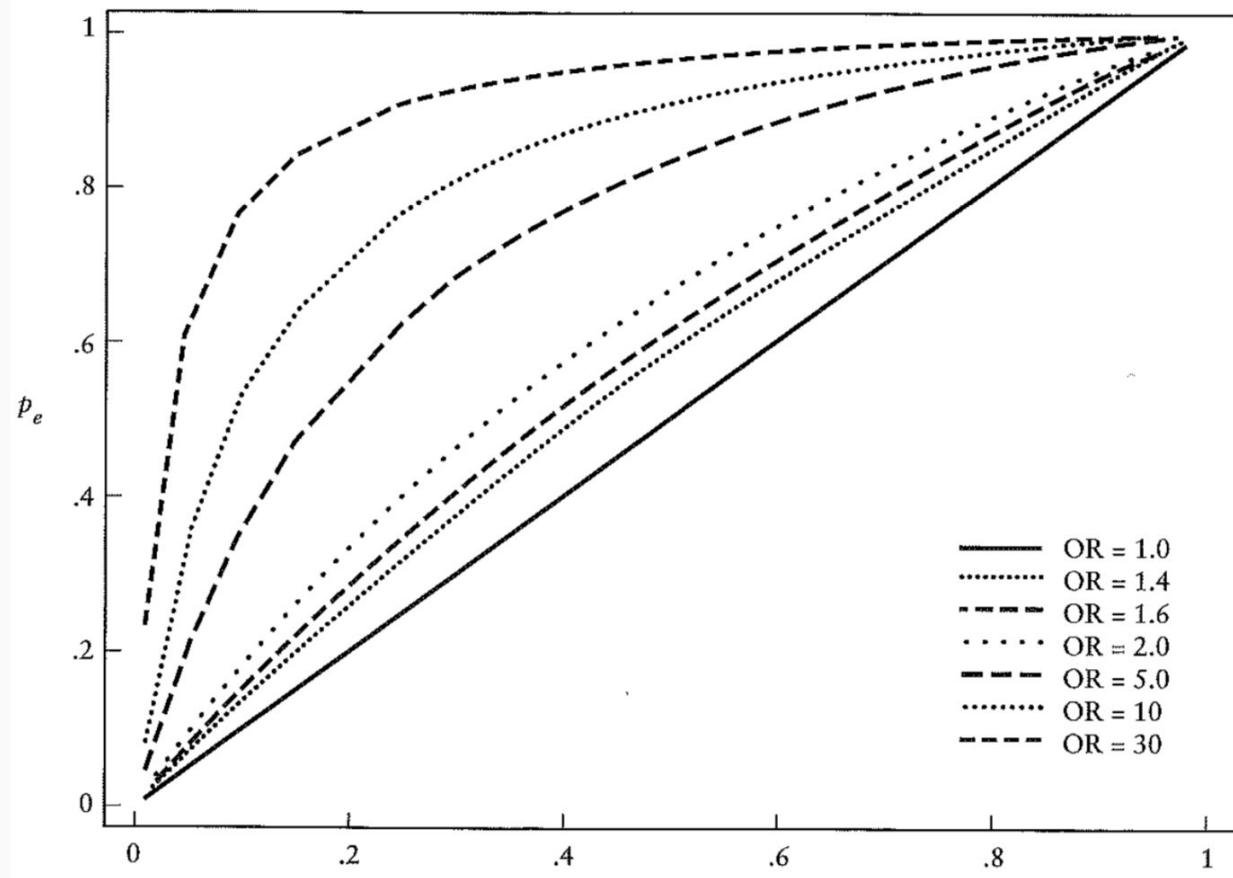
Bird OR Curve



Bird OR Curve



OR Curves



(An old) Example - House

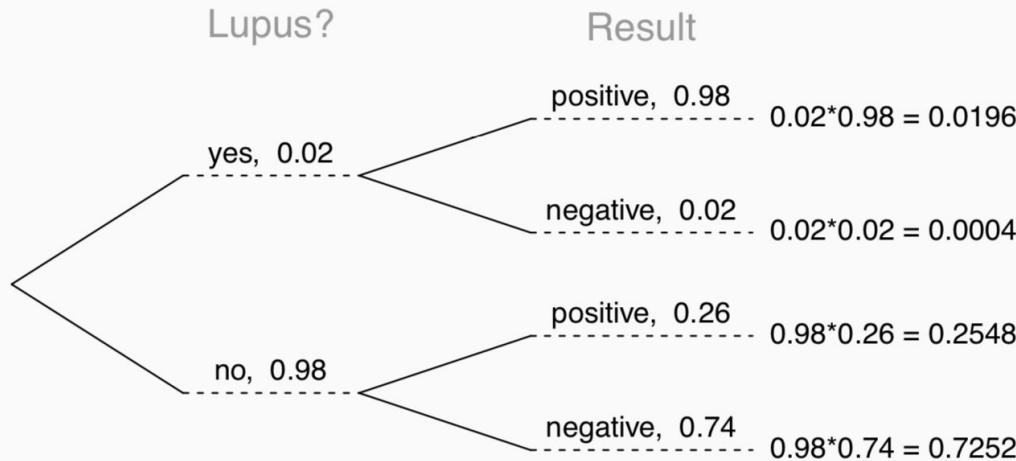
If you've ever watched the TV show House on Fox, you know that Dr. House regularly states, "It's never lupus."

Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease.

The test for lupus is very accurate if the person actually has lupus, however is very inaccurate if the person does not. More specifically, the test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease.

Is Dr. House correct even if someone tests positive for Lupus?

(An old) Example - House



$$\begin{aligned}P(\text{Lupus}|+) &= \frac{P(+, \text{Lupus})}{P(+, \text{Lupus}) + P(+, \text{No Lupus})} \\&= \frac{0.0196}{0.0196 + 0.2548} = 0.0714\end{aligned}$$

Testing for lupus

It turns out that testing for Lupus is actually quite complicated, a diagnosis usually relies on the outcome of multiple tests, often including: a complete blood count, an erythrocyte sedimentation rate, a kidney and liver assessment, a urinalysis, and or an antinuclear antibody (ANA) test.

It is important to think about what is involved in each of these tests (e.g. deciding if complete blood count is high or low) and how each of the individual tests and related decisions plays a role in the overall decision of diagnosing a patient with lupus.

Testing for lupus

At some level we can view a diagnosis as a binary decision (lupus or no lupus) that involves the complex integration of various explanatory variables.

The example does not give us any information about how a diagnosis is made, but what it does give us is just as important - the sensitivity and the specificity of the test. These values are critical for our understanding of what a positive or negative test result actually means.

Sensitivity and Specificity

Sensitivity - measures a test's ability to identify positive results.

$$P(\text{Test} + \mid \text{Condition} +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a test's ability to identify negative results.

$$P(\text{Test} - \mid \text{Condition} -) = P(-|\text{no lupus}) = 0.74$$

Sensitivity and Specificity

Sensitivity - measures a test's ability to identify positive results.

$$P(\text{Test} + \mid \text{Condition} +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a test's ability to identify negative results.

$$P(\text{Test} - \mid \text{Condition} -) = P(-|\text{no lupus}) = 0.74$$

It is illustrative to think about the extreme cases - what is the sensitivity and specificity of a test that always returns a positive result? What about a test that always returns a negative result?

Sensitivity and Specificity (cont.)

		Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)	
Test Negative	False Negative (Type II error)	True Negative	

$$\text{Sensitivity} = P(\text{Test +} \mid \text{Condition +}) = TP / (TP + FN)$$

$$\text{Specificity} = P(\text{Test -} \mid \text{Condition -}) = TN / (FP + TN)$$

$$\text{False negative rate } (\beta) = P(\text{Test -} \mid \text{Condition +}) = FN / (TP + FN)$$

$$\text{False positive rate } (\alpha) = P(\text{Test +} \mid \text{Condition -}) = FP / (FP + TN)$$

$$\text{Sensitivity} = 1 - \text{False negative rate} = \text{Power}$$

$$\text{Specificity} = 1 - \text{False positive rate}$$

So what?

Clearly it is important to know the Sensitivity and Specificity of test (and or the false positive and false negative rates). Along with the incidence of the disease (e.g. $P(\text{lupus})$) these values are necessary to calculate important quantities like $P(\text{lupus}|+)$.

Additionally, our brief foray into power analysis before the first midterm should also give you an idea about the trade offs that are inherent in minimizing false positive and false negative rates (increasing power required either increasing α or n).

How should we use this information when we are trying to come up with a decision?

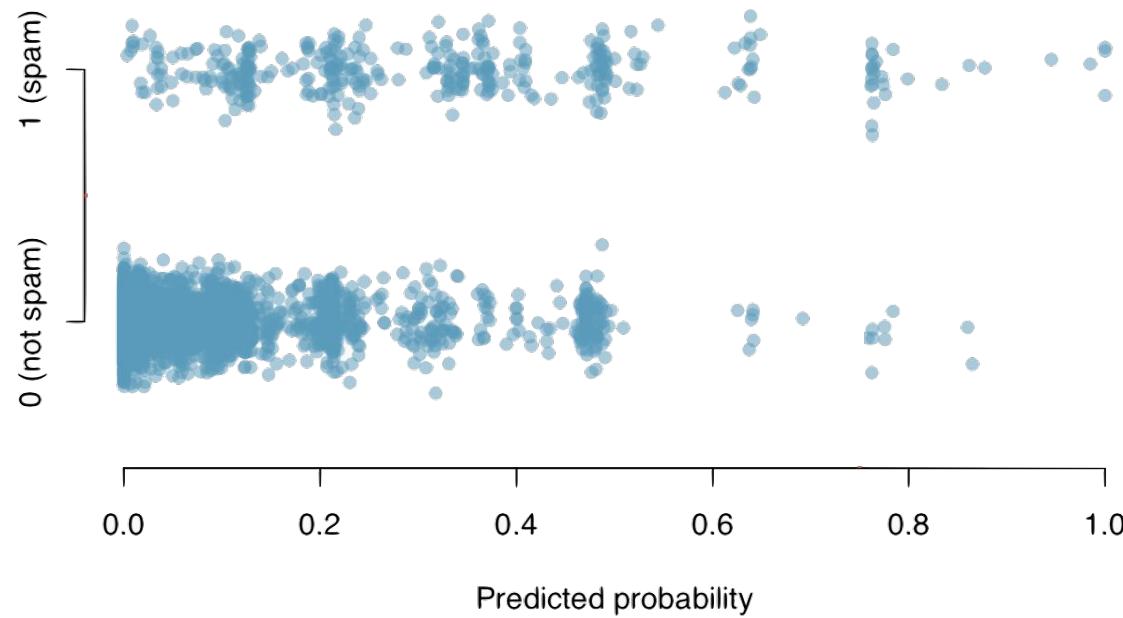
Back to Spam

In lab this week, we examined a data set of emails where we were interesting in identifying the spam messages. We examined different logistic regression models to evaluate how different predictors influenced the probability of a message being spam.

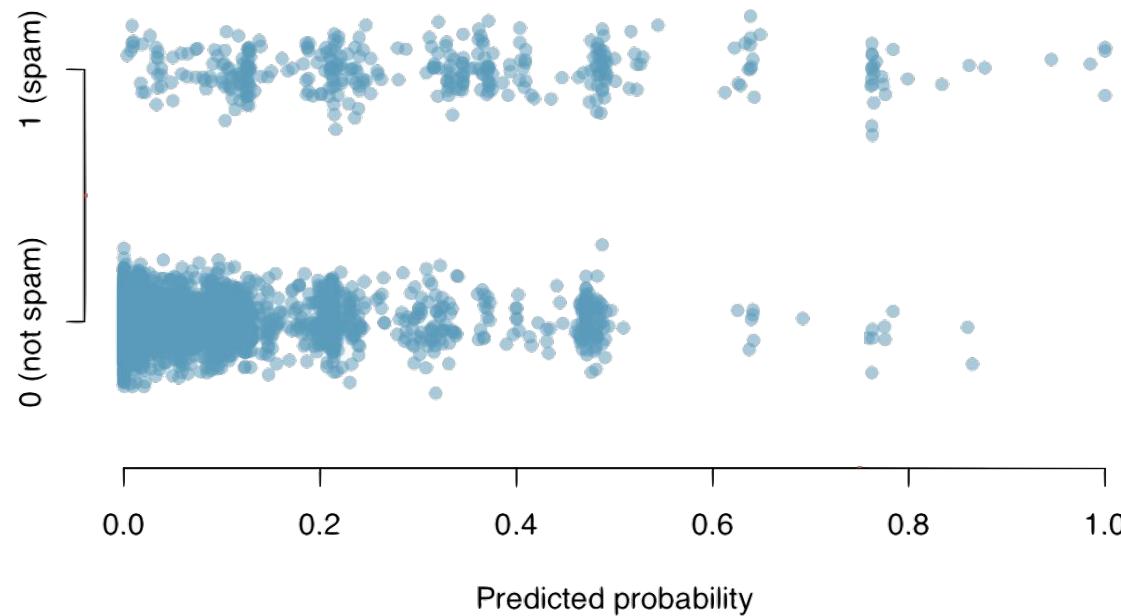
These models can also be used to assign probabilities to incoming messages (this is equivalent to prediction in the case of SLR / MLR). However, if we were designing a spam filter this would only be half of the battle, we would also need to use these probabilities to make a decision about which emails get flagged as spam.

While not the only possible solution, we will consider a simple approach where we choose a threshold probability and any email that exceeds that probability is flagged as spam.

Picking a threshold

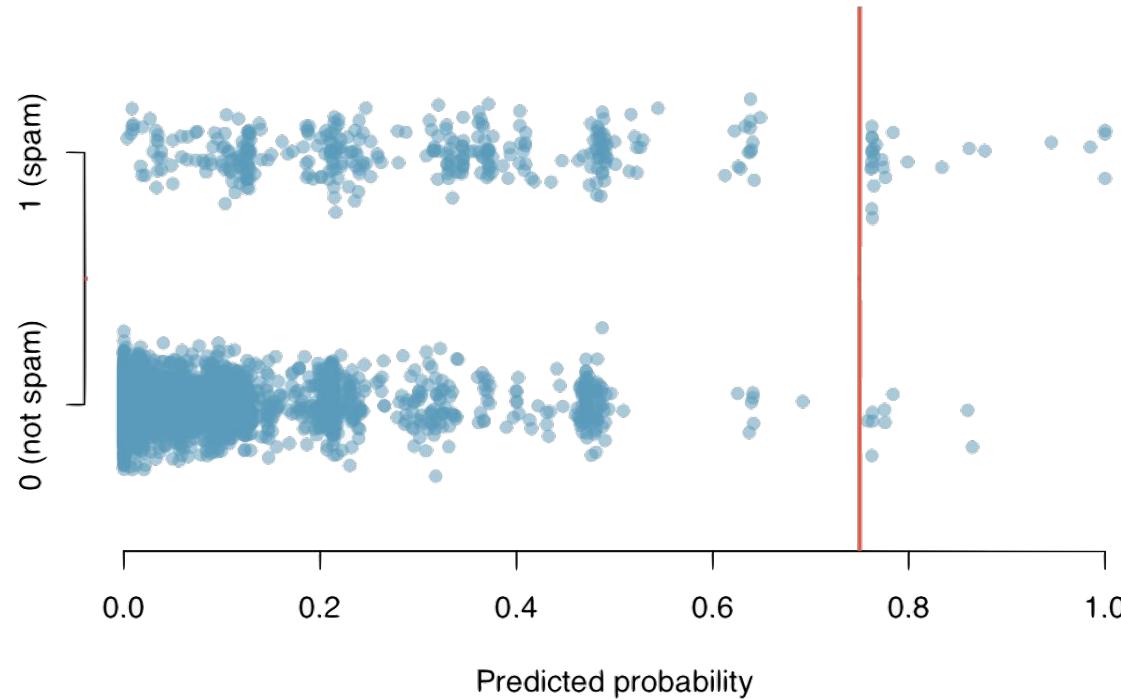


Picking a threshold



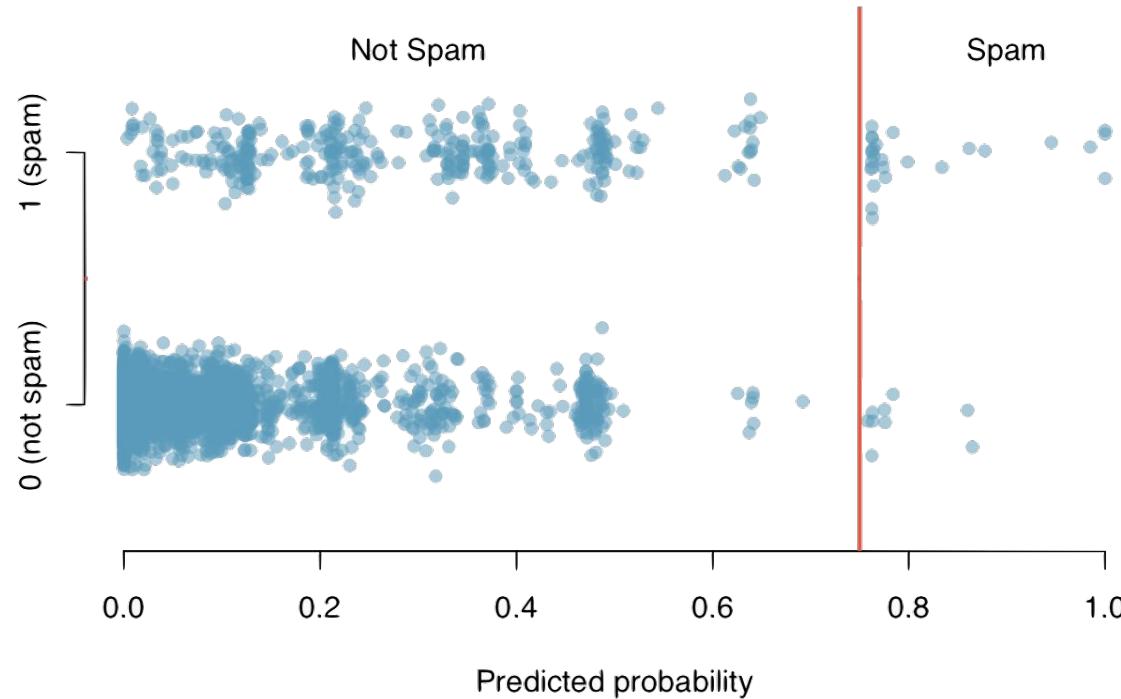
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



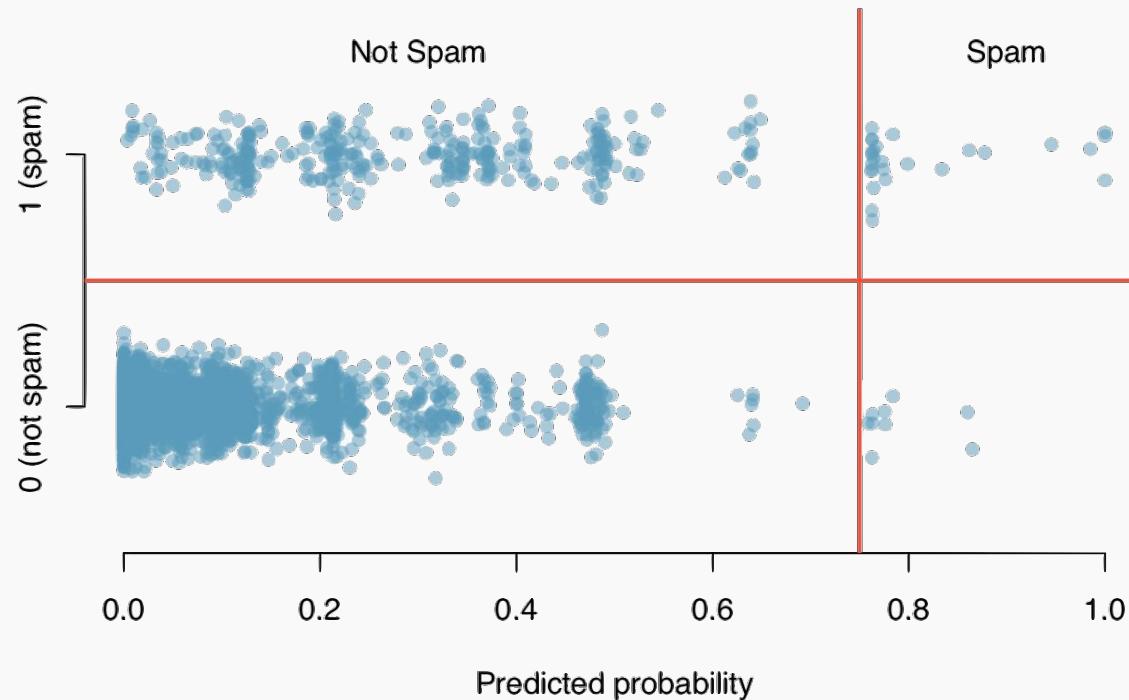
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



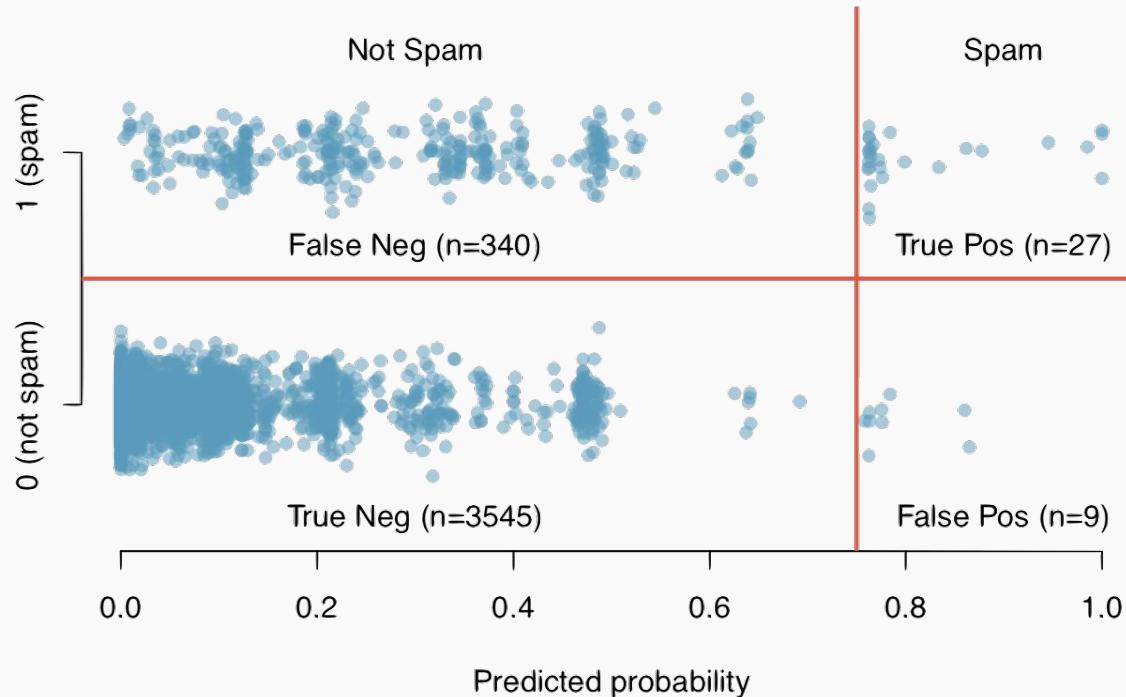
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



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

Picking a threshold



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

Consequences of picking a threshold

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

$$\begin{array}{ll} \text{FN} = 340 & \text{TP} = 27 \\ \text{TN} = 3545 & \text{FP} = 9 \end{array}$$

Consequences of picking a threshold

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

$$\begin{array}{ll} \text{FN} = 340 & \text{TP} = 27 \\ \text{TN} = 3545 & \text{FP} = 9 \end{array}$$

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

Consequences of picking a threshold

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

$$FN = 340$$

$$TP = 27$$

$$TN = 3545$$

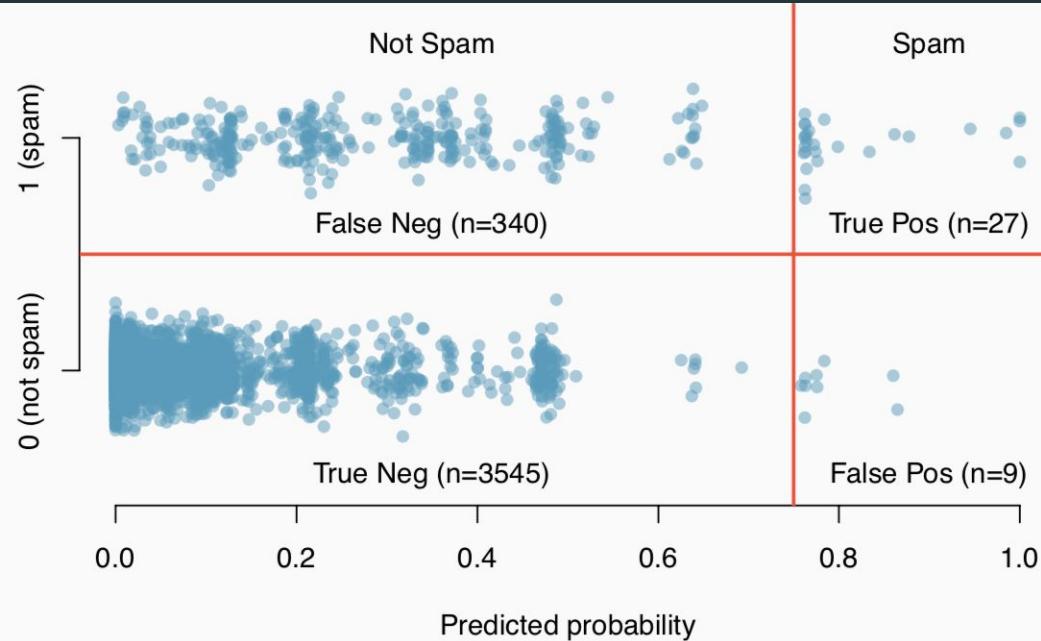
$$FP = 9$$

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

$$\text{Sensitivity} = TP/(TP + FN) = 27/(27 + 340) = 0.073$$

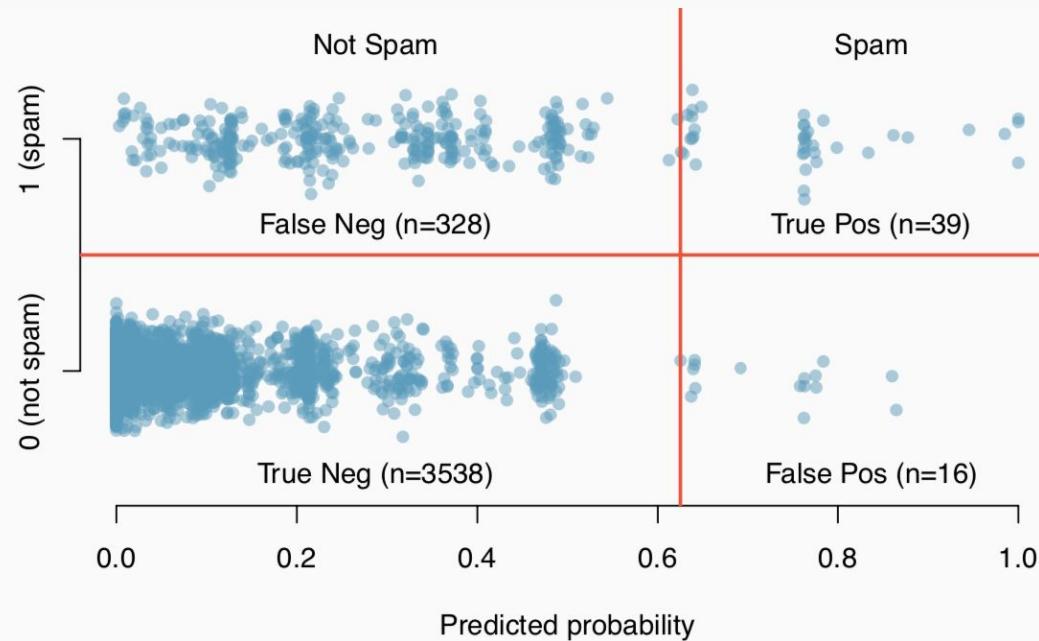
$$\text{Specificity} = TN/(FP + TN) = 3545/(9 + 3545) = 0.997$$

Trying other thresholds



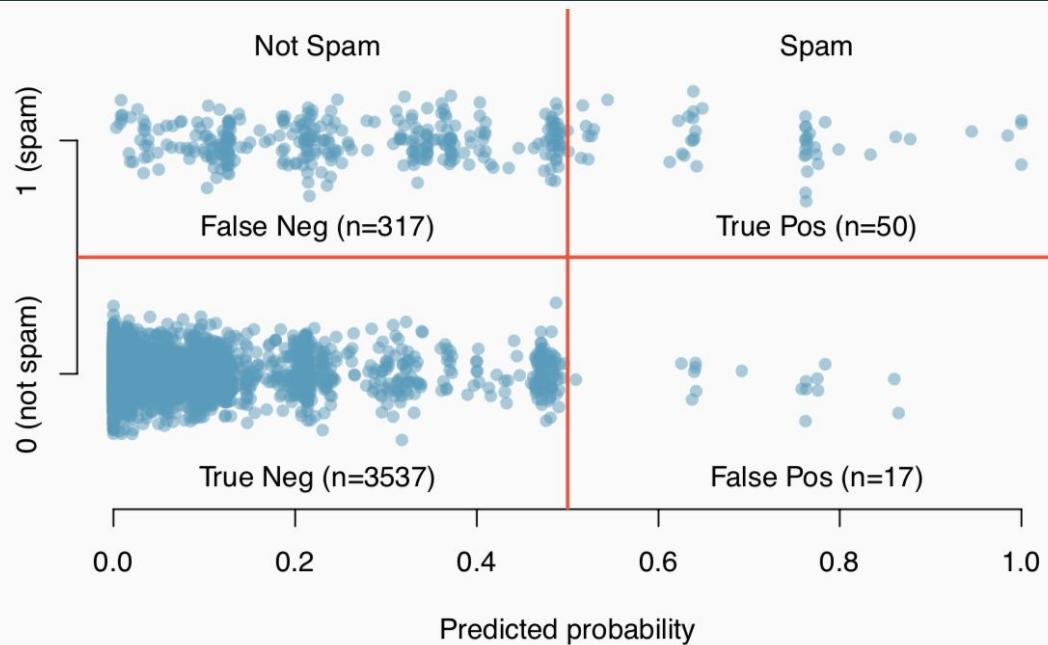
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Trying other thresholds



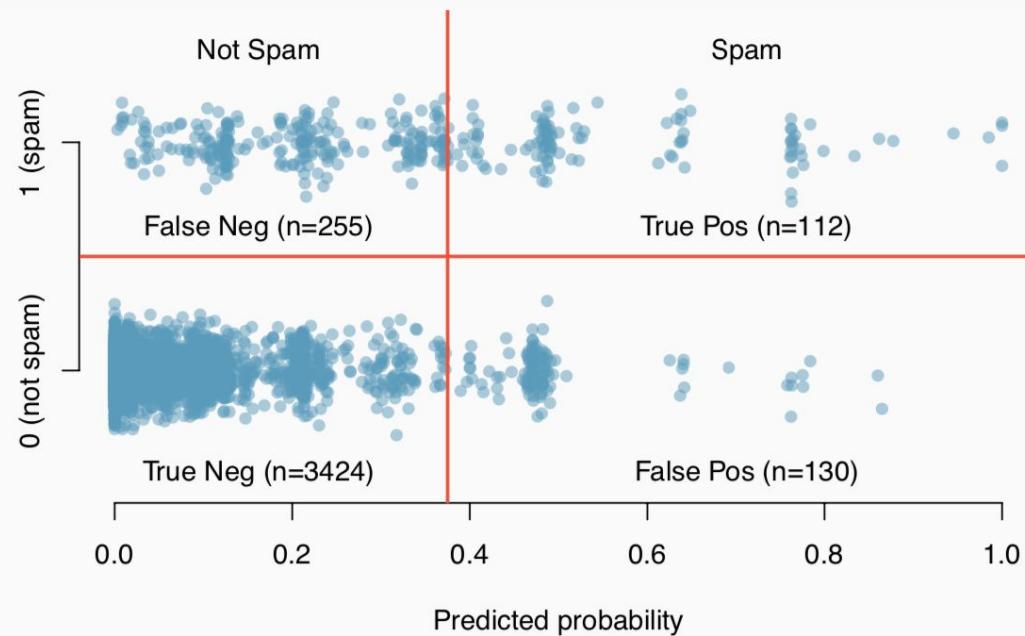
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



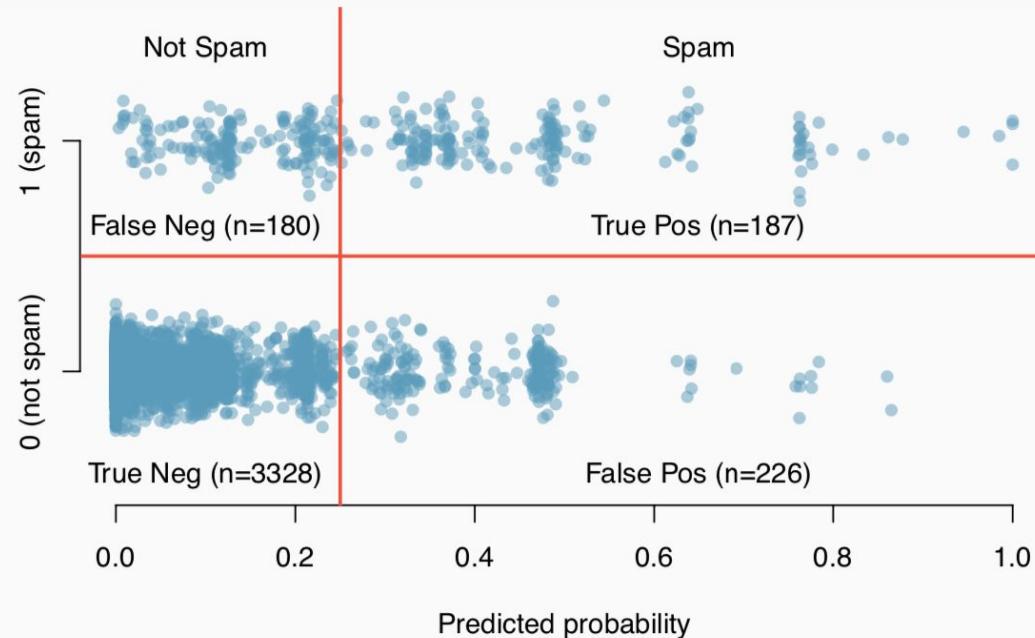
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Trying other thresholds



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	

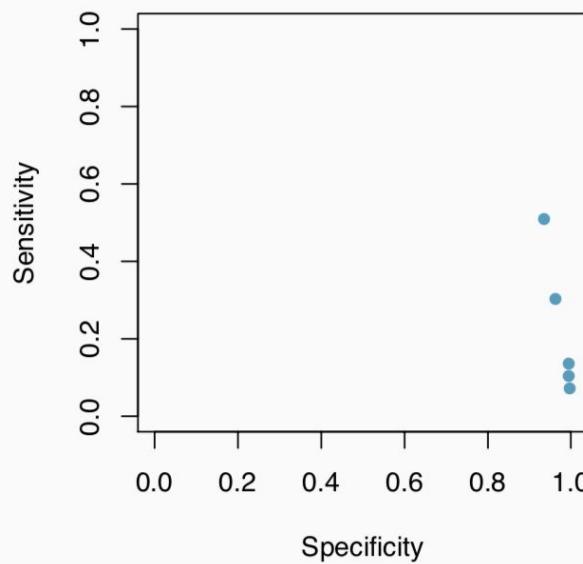
Trying other thresholds



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

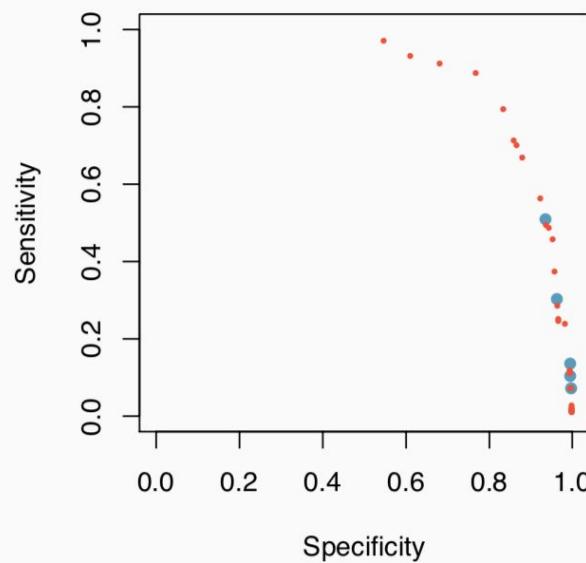
Relationship between Sensitivity and Specificity

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



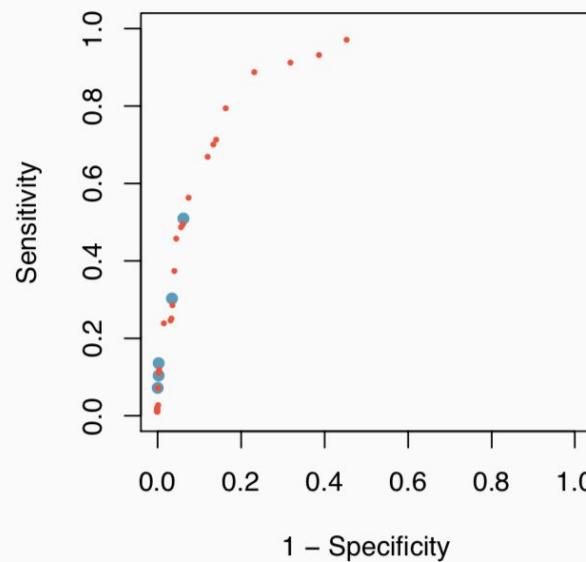
Relationship between Sensitivity and Specificity

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

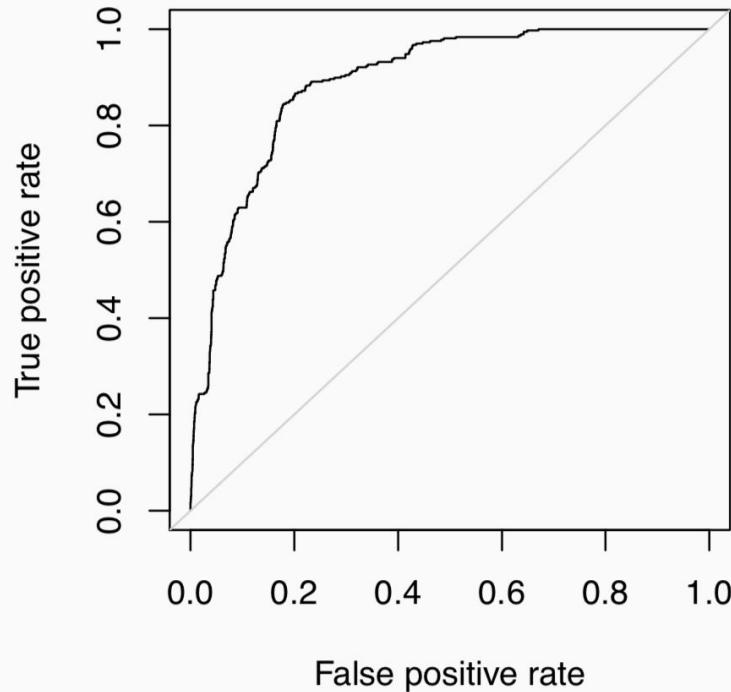


Relationship between Sensitivity and Specificity

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



Receiver operating characteristic (ROC) curve



Receiver operating characteristic (ROC) curve (cont.)

Why do we care about ROC curves?

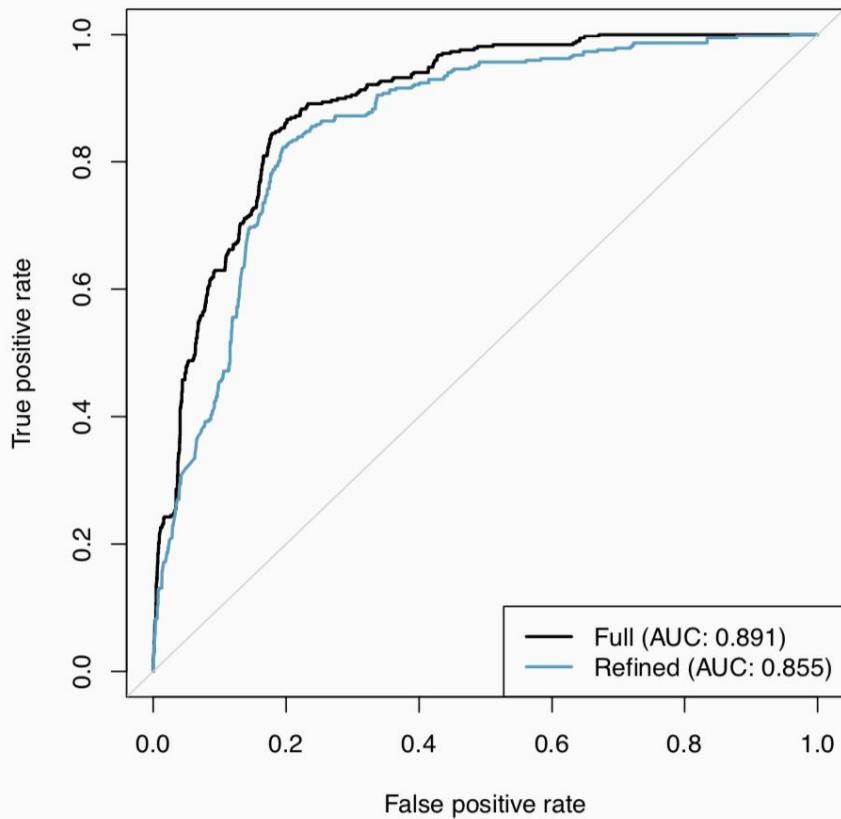
- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Refining the Spam model

```
g_refined = glm(spam ~ to_multiple+cc+image+attach+winner  
+password+line_breaks+format+re_subj  
+urgent_subj+exclaim_mess,  
data=email, family=binomial)  
summary(g_refined)
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multipleyes	-2.7368	0.3156	-8.67	0.0000
ccyes	-0.5358	0.3143	-1.71	0.0882
imageyes	-1.8585	0.7701	-2.41	0.0158
attachyes	1.2002	0.2391	5.02	0.0000
winneryes	2.0433	0.3528	5.79	0.0000
passwordyes	-1.5618	0.5354	-2.92	0.0035
line_breaks	-0.0031	0.0005	-6.33	0.0000
formatPlain	1.0130	0.1380	7.34	0.0000
re_subjyes	-2.9935	0.3778	-7.92	0.0000
urgent_subjyes	3.8830	1.0054	3.86	0.0001
exclaim_mess	0.0093	0.0016	5.71	0.0000

Comparing models



Utility Functions

There are many other reasonable quantitative approaches we can use to decide on what is the “best” threshold.

If you've taken an economics course you have probably heard of the idea of utility functions, we can assign costs and benefits to each of the possible outcomes and use those to calculate a utility for each circumstance.

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each outcome.

Outcome	Utility
True Positive	
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each outcome.

Outcome	Utility
True Positive	1
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each outcome.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each outcome.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each outcome.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each outcome.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

$$U(p) = TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p)$$

Utility for the 0.75 threshold

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

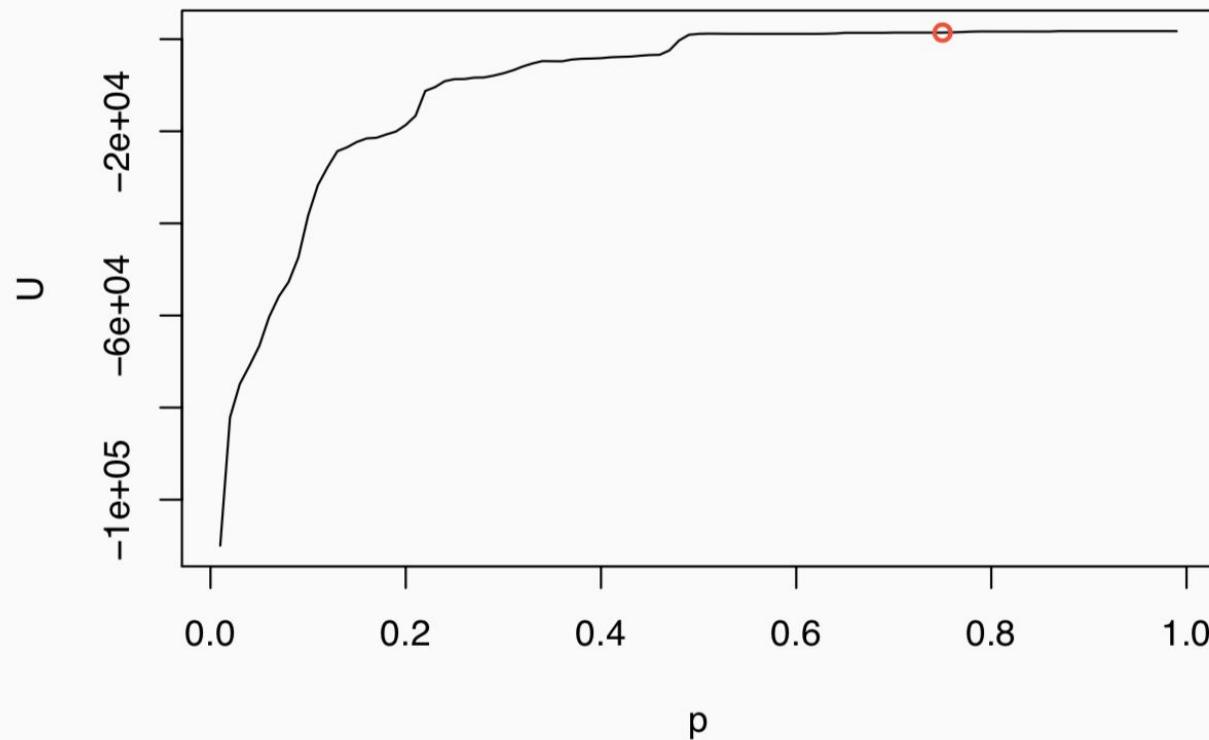
$$FN = 340 \quad TP = 27$$

$$TN = 3545 \quad FP = 9$$

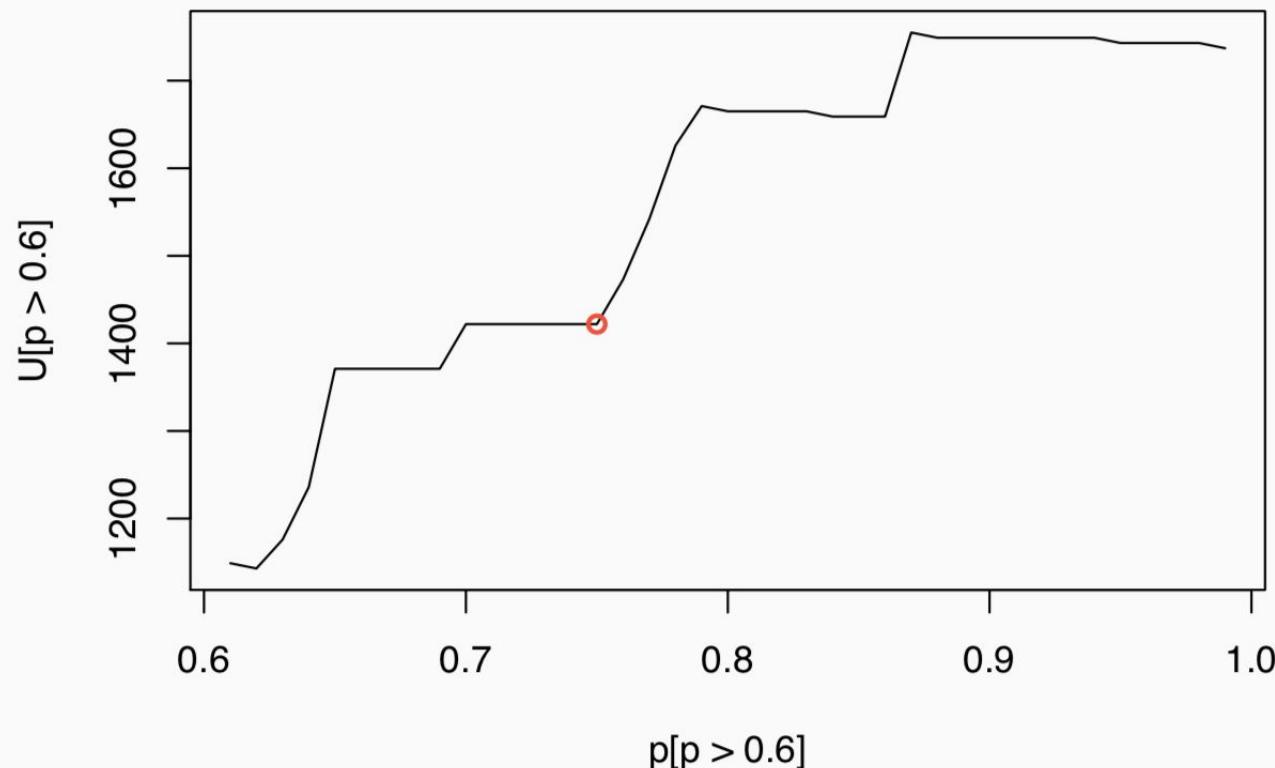
$$\begin{aligned} U(p) &= TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p) \\ &= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422 \end{aligned}$$

Not useful by itself, but allows us to compare with other thresholds.

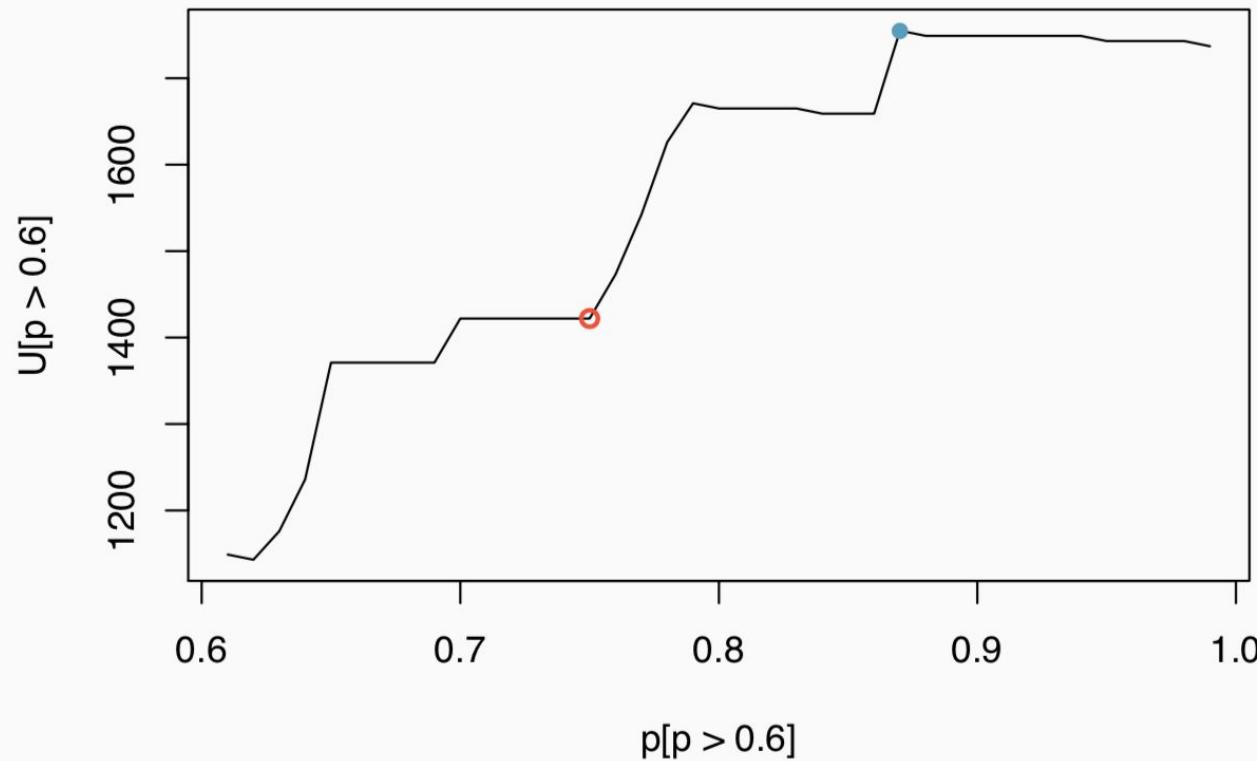
Utility curve



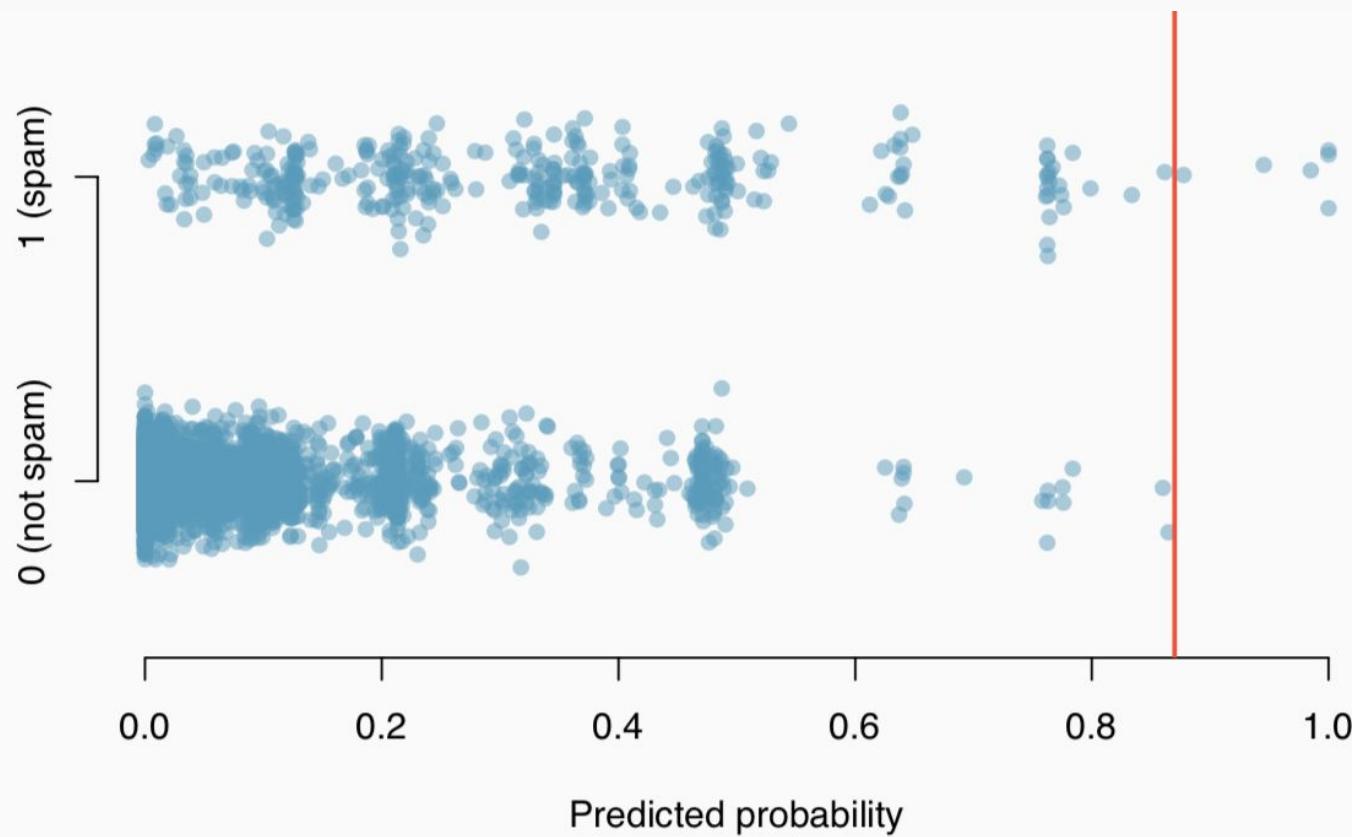
Utility curve (zoom)



Utility curve (zoom)



Maximum Utility



Find more resources at openintro.org/os, including

- Slides
- Videos
- Statistical Software Labs
- Discussion Forums (free support for students and teachers)
- Learning Objectives

Teachers only content is also available for [Verified Teachers](#), including

- Exercise solutions
- Sample exams
- Ability to request a free desk copy for a course
- Statistics Teachers email group

Questions? [Contact us](#).