Statistical Inference

Logistic Regression

Behnam Bahrak

Spring 2020

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.

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

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

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

p-value $= P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10) = 2 \times 0.0178 = 0.0359$

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

- 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:

$$e^{CI} = (e^{-0.1513}, e^{-0.0051})$$

همه چی
بر اساس Log
حساب میشه

Example: Birdkeeping and Lung Cancer

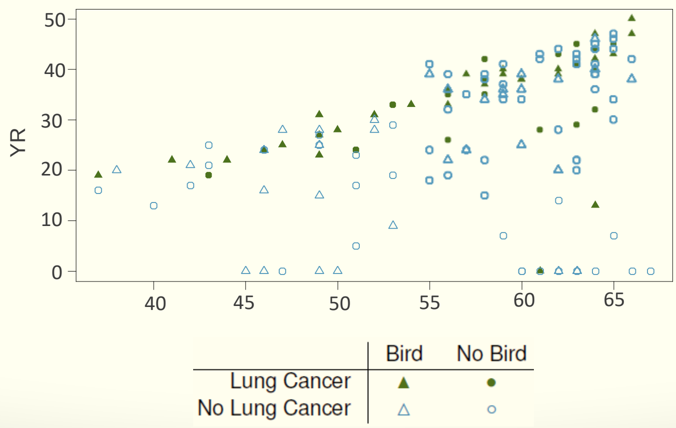
- A 1972 - 1981 health survey in Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer.
- To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in Hague (population 450,000).
- 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.

Birdkeeping and Lung Cancer - Data

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37	19	12
2	LungCancer	Male	Low	Bird	41	22	15
3	LungCancer	Male	High	NoBird	43	19	15
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65	7	2

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)

Birdkeeping and Lung Cancer



Interpretation

R

```
> summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FM:Female	0.5613	0.5312	1.06	0.2907
SS:High	0.1054	0.4688	0.22	0.8221
BK:Bird	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 $e^{1.3626} = 3.91$
  - The odds ratio of getting lung cancer for an additional year of smoking is $e^{0.0729} = 1.08$

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.

داشتن بیماری

پرنده داشتن

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

نداشتن بیماری

پرنده نداشتن

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

این رو RR میده

Example

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

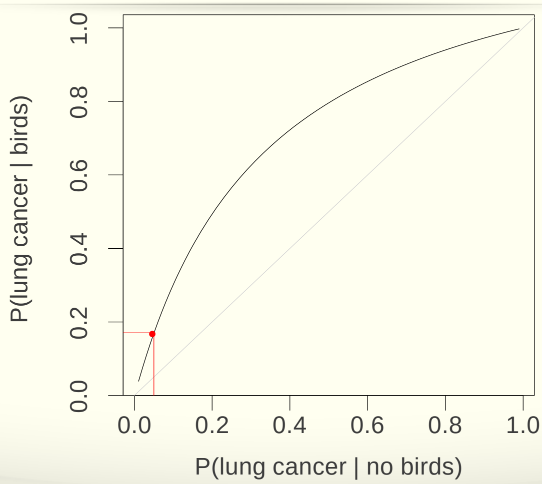
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1-P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1-P(\text{lung cancer}|\text{no birds})]}$$

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

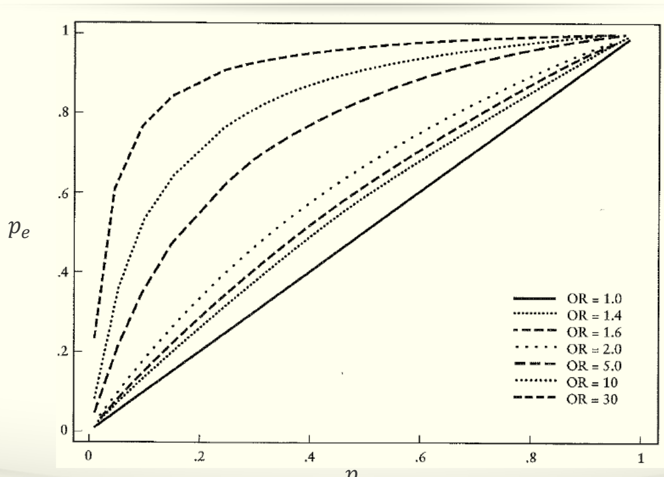
$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

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

Bird Odds Ratio (OR) Curve



OR Curves



Maximum Likelihood

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- We use maximum likelihood method to estimate β_0 and β_1:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

- We pick β_0 and β_1 to maximize the likelihood of the observed data: $l(\beta_0, \beta_1)$

Example - House

- If you've ever watched the TV show House M.D., 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 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?

Example - House

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

+| بیمار

-| سالم

Testing for Lupus

- It turns out that testing for Lupus is actually quite complicated, a diagnosis usually relies on the outcome of multiple tests.
- It is important to think about what is involved in each of these tests and how each of the individual tests and related decisions plays a role in the overall decision of diagnosing a patient with Lupus.
- The example gives us 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 (Recall or True Positive)- measures a tests ability to identify positive results.

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

- Specificity (True Negative)- measures a tests ability to identify negative results.

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

- 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

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type 1 error)
Test Negative	False Negative (Type 2 error)	True Negative

Sensitivity = P(Test + | Condition +) = TP/(TP + FN)
Specificity = P(Test - | Condition -) = TN/(FP + TN)
False negative rate (β) = P(Test - | Condition +) = FN/(TP + FN)
False positive rate (α) = P(Test + | Condition -) = FP/(FP + TN)

Sensitivity = 1 - False negative rate = Power
Specificity = 1 - 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(lupus)) these values are necessary to calculate important quantities like P(lupus|+).

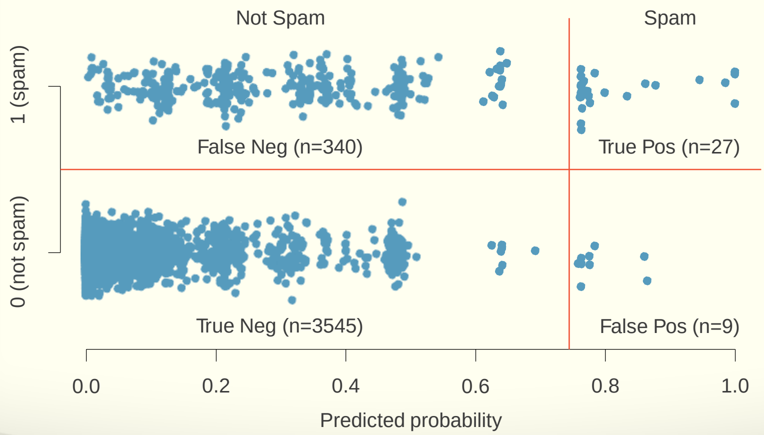
- Additionally, our brief foray into power analysis should also give you an idea about the trade offs that are inherent in minimizing false positive and false negative rates (increasing power required either increasing α or n).

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

Spam

- We examine a data set of emails where we are interested in identifying the spam messages.
- We can use 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 emails.
- We also need to use these probabilities to make a decision about which emails get flagged as spam.
- While not the only possible solution, we consider a simple approach where we choose a threshold probability and any email that exceeds that probability is flagged as spam.

Picking a Threshold



Consequences of Picking a Threshold

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

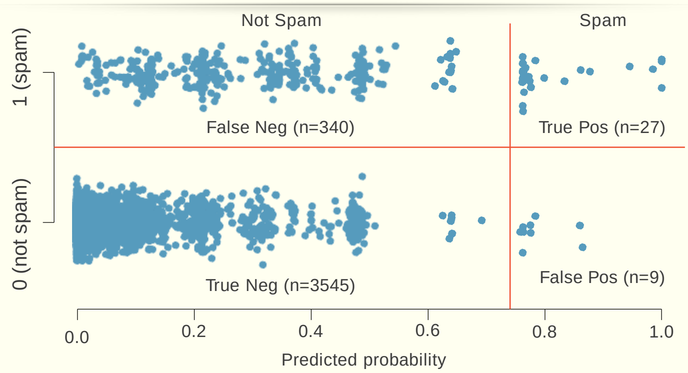
$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

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

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

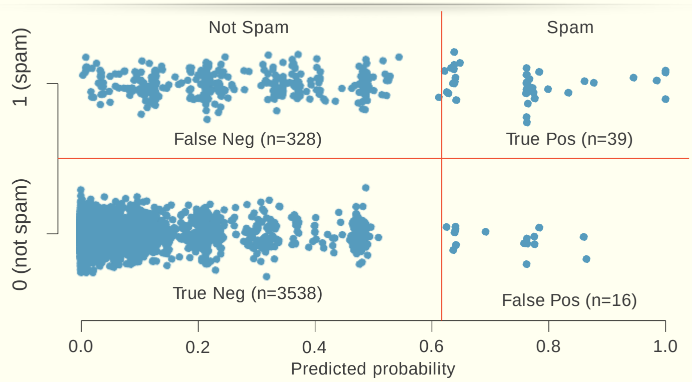
$$\text{Specificity} = TN/(FP + TN) = 3545/(9 + 3545) = 0.997$$

Trying other thresholds



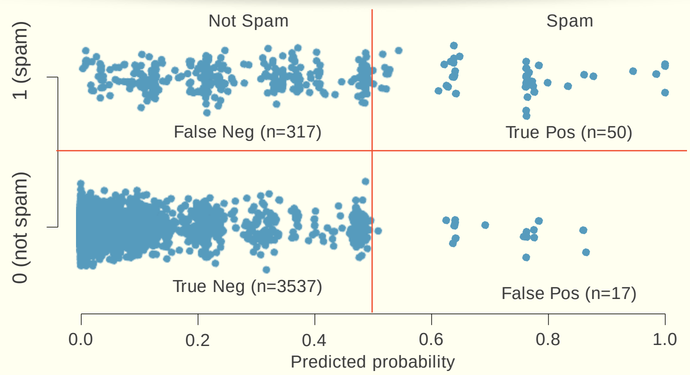
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Trying other thresholds



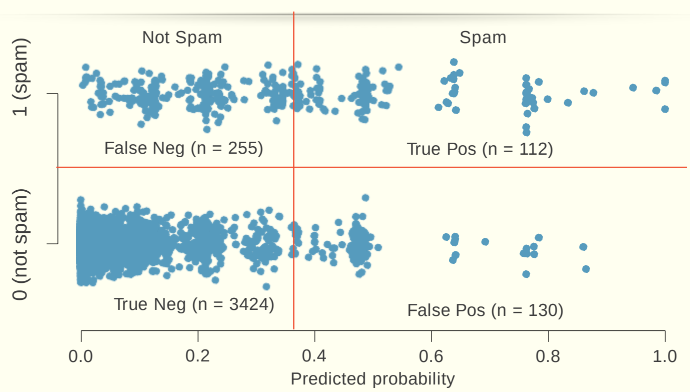
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



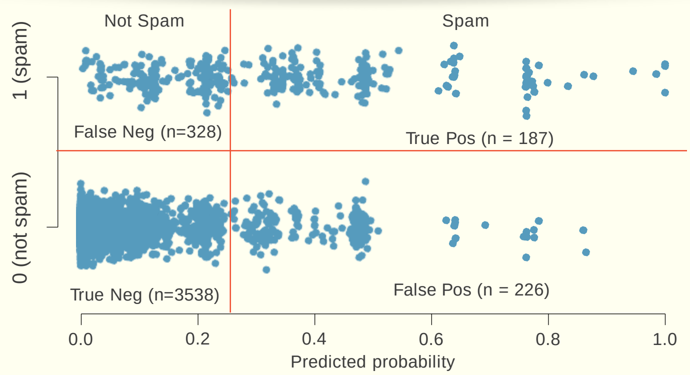
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

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	

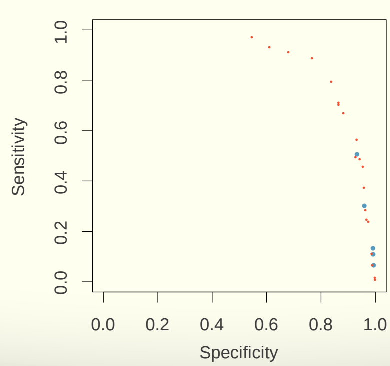
Trying other thresholds



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

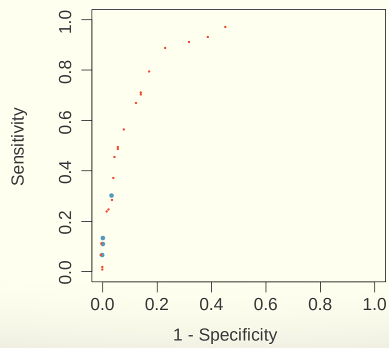
Relationship btw Sensitivity & Specificity

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

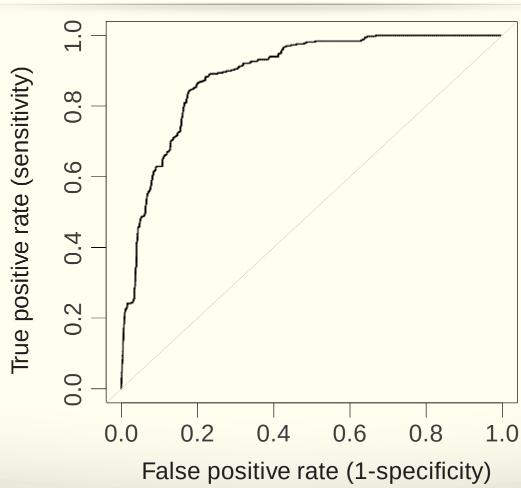


Relationship btw Sensitivity & Specificity

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

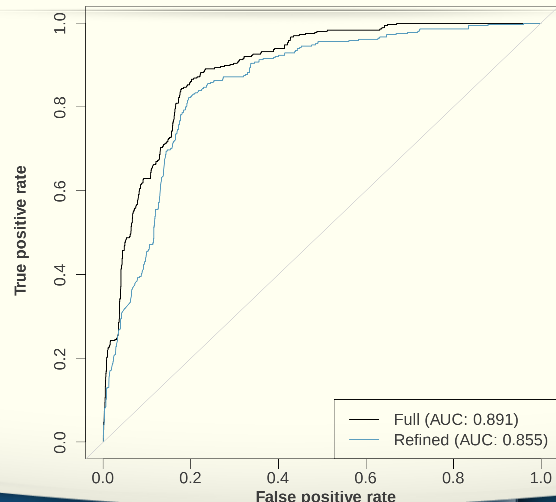
- Why do we care about ROC curves?
  - Shows the trade off in sensitivity and specificity for all possible thresholds.
  - Straight forward to compare performance vs. chance.
  - Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Refining the Spam Model

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multiple:yes	-2.7368	0.3156	-8.67	0.0000
cc:yes	-0.5358	0.3143	-1.71	0.0882
image:yes	-1.8585	0.7701	-2.41	0.0158
attach:yes	1.2002	0.2391	5.02	0.0000
winner:yes	2.0433	0.3528	5.79	0.0000
password:yes	-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_subj:yes	-2.9935	0.3778	-7.92	0.0000
urgent_subj:yes	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 out.

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:

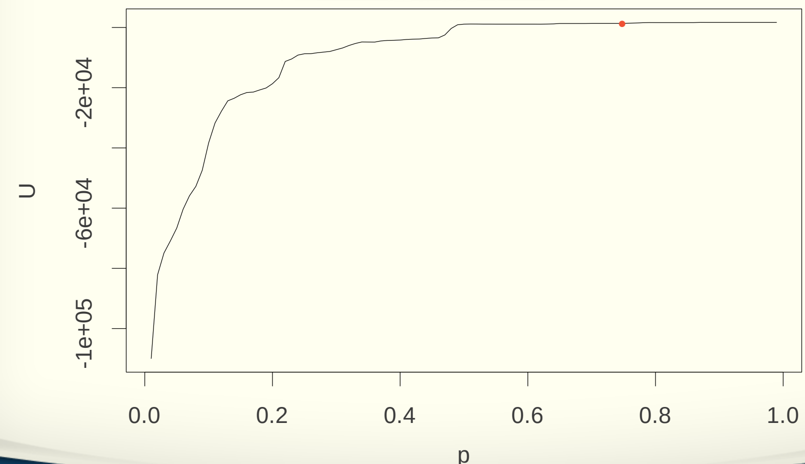
$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

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

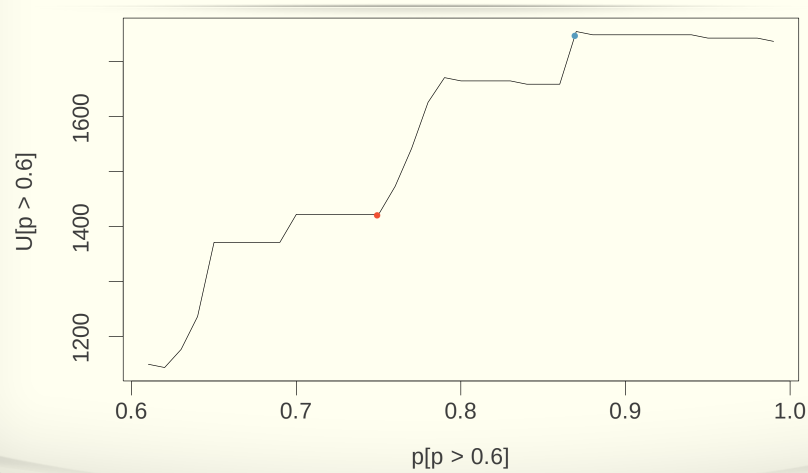
$$= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422$$

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

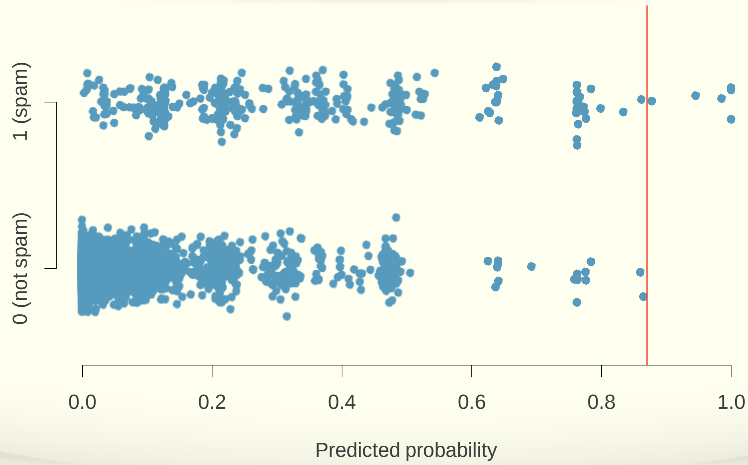
Utility Curve



Utility curve (zoom)



Maximum Utility



Logistic Regression Conditions

- There are three key conditions for fitting a logistic regression model:

  1. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.

  2. Each outcome Y_i is independent of the other outcomes.

  3. There should be little or no multicollinearity among the explanatory variables.