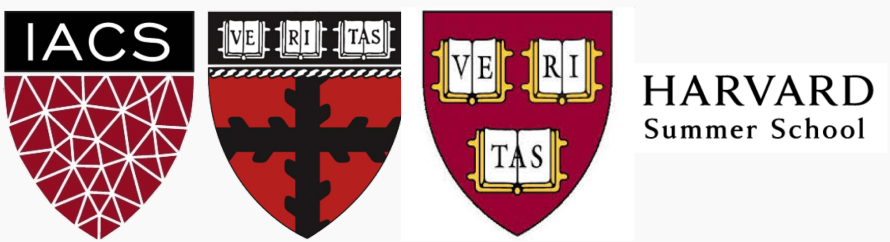
the Y-variable is categorical

Lecture #6: Logistic Regression

CS-S109A: Introduction to Data Science

Kevin Rader



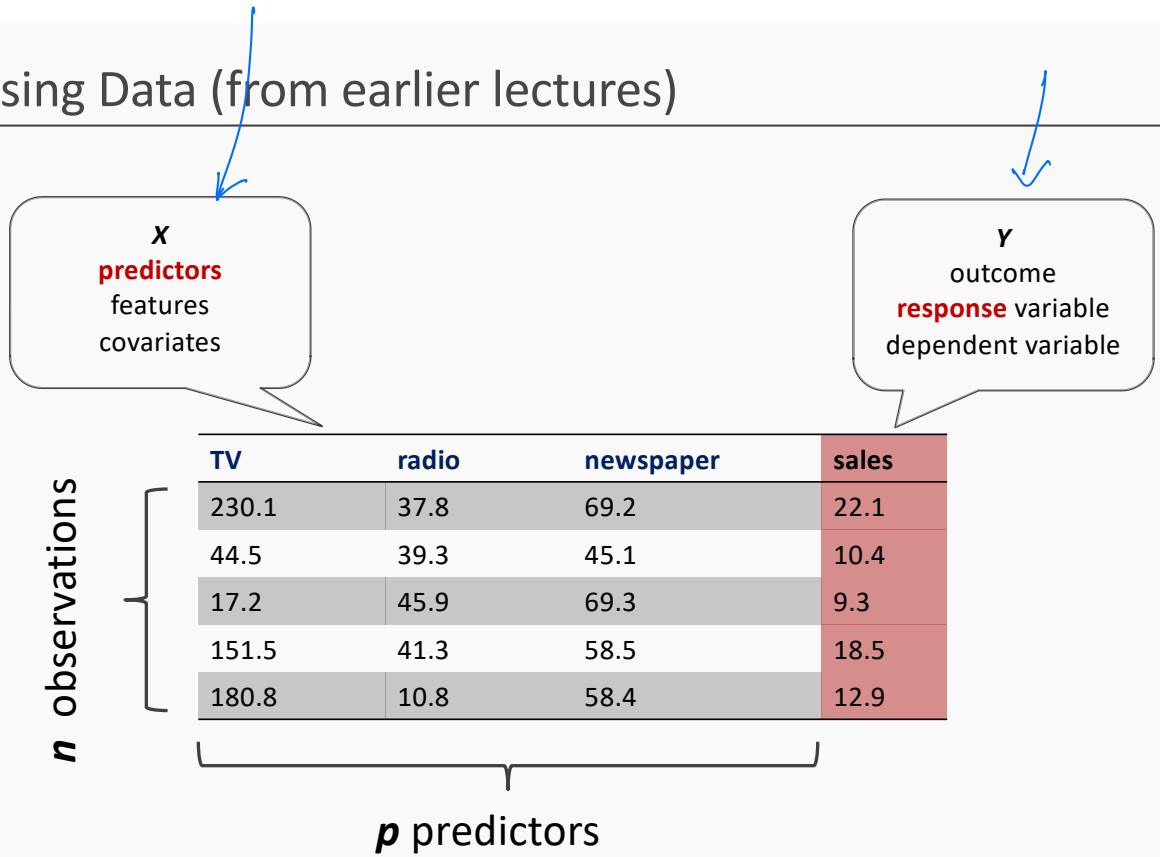
Lecture Outline

- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
  - Estimating the Simple Logistic Model
  - Classification using the Logistic Model
  - Multiple Logistic Regression
  - Extending the Logistic Model
- Classification Boundaries
- Regularization in Logistic Regression
- Multinomial Logistic Regression
- Bayes Theorem and Misclassification Rates
- ROC Curves

1st half

2nd half

Advertising Data (from earlier lectures)



Heart Data

response variable ***Y*** is Yes/No

Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
63	1	typical	145	233	1	2	150	0	2.3	3	0.0	fixed	No
67	1	asymptomatic	160	286	0	2	108	1	1.5	2	3.0	normal	Yes
67	1	asymptomatic	120	229	0	2	129	1	2.6	2	2.0	reversable	Yes
37	1	nonanginal	130	250	0	0	187	0	3.5	3	0.0	normal	No
41	0	nontypical	130	204	0	2	172	0	1.4	1	0.0	normal	No

Heart Data

These data contain a binary outcome HD for 303 patients who presented with chest pain. An outcome value of:

- ***Yes*** indicates the presence of heart disease based on an angiographic test,
- ***No*** means no heart disease.

There are 13 predictors including:

- Age
- Sex (0 for women, 1 for men)
- Chol (a cholesterol measurement),
- MaxHR
- RestBP (systolic)

and other heart and lung function measurements.

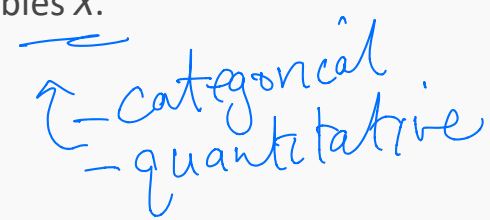
Classification

Classification

Up to this point, the methods we have seen have centered around modeling and the prediction of a **quantitative** response variable (ex, number of taxi pickups, number of bike rentals, etc). Linear **regression** (and Ridge, LASSO, etc) perform well under these situations

When the response variable is **categorical**, then the problem is no longer called a regression problem but is instead labeled as a **classification problem**.

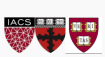
The goal is to attempt to classify each observation into a category (aka, class or cluster) defined by *Y*, based on a set of predictor variables *X*.



Typical Classification Examples

The motivating examples for this lecture(s), homework, and coming labs are based [mostly] on medical data sets. Classification problems are common in this domain:

- Trying to determine where to set the *cut-off* for some diagnostic test (pregnancy tests, prostate or breast cancer screening tests, etc...)
- Trying to determine if cancer has gone into remission based on treatment and various other indicators
- Trying to classify patients into types or classes of disease based on various genomic markers



Why not Linear Regression?

Simple Classification Example

Given a dataset:

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$$

where the y are categorical (sometimes referred to as ***qualitative***), we would like to be able to predict which category y takes on given x.

A categorical variable y could be encoded to be quantitative. For example, if y represents concentration of Harvard undergrads, then y could take on the values:

$$y = \begin{cases} 1 & if \text{ Computer Science (CS)} \\ 2 & if \text{ Statistics} \\ 3 & \text{otherwise} \end{cases}.$$

Scale or ordering of categories affects the interpretation

Linear regression **does not work well**, or is not appropriate at all, in this setting.

Simple Classification Example (cont.)

A linear regression could be used to predict y from $\boldsymbol{x}$. What would be wrong with such a model?

The model would imply a specific ordering of the outcome, and would treat a one-unit change in y equivalent. The jump from $y = 1$ to $y = 2$ (CS to Statistics) should not be interpreted as the same as a jump from $y = 2$ to $y = 3$ (Statistics to everyone else).

Similarly, the response variable could be reordered such that $y = 1$ represents Statistics and $y = 2$ represents CS, and then the model estimates and predictions would be fundamentally different.

If the categorical response variable was ***ordinal*** (had a natural ordering, like class year, Freshman, Sophomore, etc.), then a linear regression model would make some sense but is still not ideal.

Even Simpler Classification Problem: Binary Response

The simplest form of classification is when the response variable y has only two categories, and then an ordering of the categories is natural. For example, an upperclassmen Harvard student could be categorized as (note, the $y = 0$ category is a "catch-all" so it would involve both River House students and those who live in other situations: off campus, etc):

$$y = \begin{cases} 1 & if \text{ lives in the Quad} \\ 0 & \text{otherwise} \end{cases}.$$

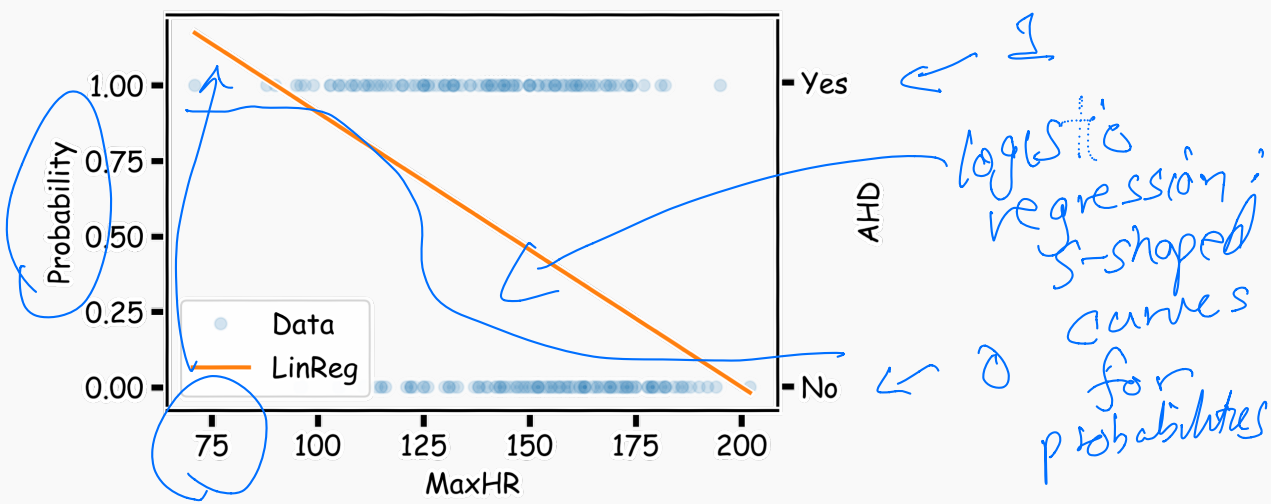
Linear regression could be used to predict y directly from a set of covariates (like sex, whether an athlete or not, concentration, GPA, etc.), and if $\hat{y} \geq 0.5$, we could predict the student lives in the Quad and predict other houses if $\hat{y} < 0.5$.

$\hat{y} > 1$ or $\hat{y} < 0$

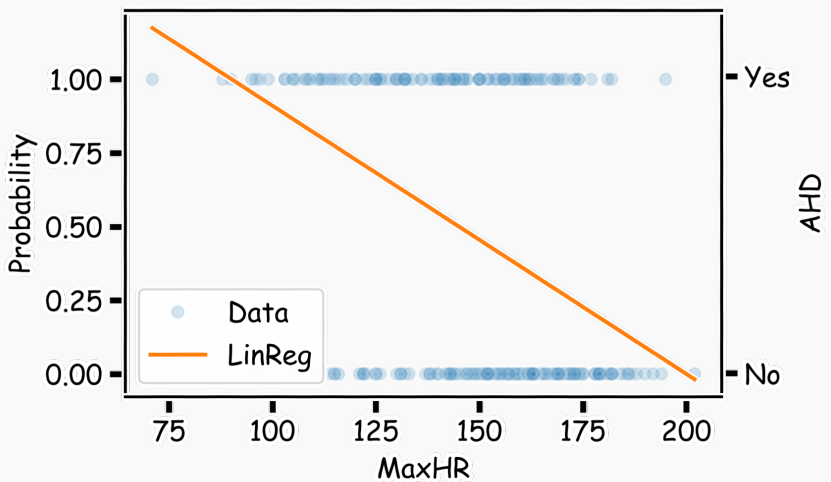
Even Simpler Classification Problem: Binary Response (cont)

What could go wrong with this linear regression model?

.



Even Simpler Classification Problem: Binary Response (cont)



The main issue is you could get non-sensical values for y. Since this is modeling $P(y = 1)$, values for $\hat{y}$ below 0 and above 1 would be at odds with the natural measure for y. Linear regression can lead to this issue.

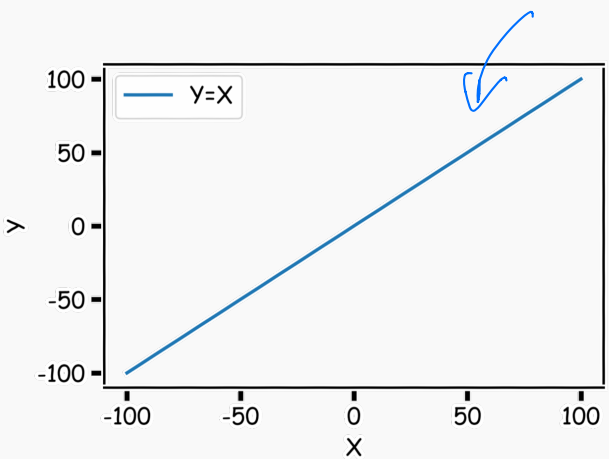
Binary Response & Logistic Regression

Pavlos Game #45

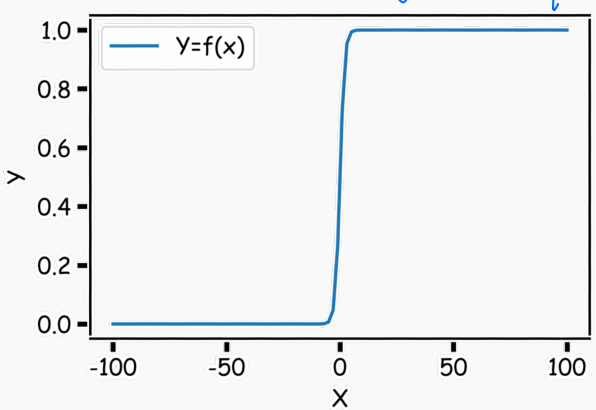
Think of a function that would do this for us

logistic function:

$$Y = \frac{1}{1+e^{(-x)}}$$



$$Y = f(x)$$



Logistic Regression

Logistic Regression addresses the problem of estimating a probability, $P(y = 1)$, to be outside the range of $[0,1]$. The logistic regression model uses a function, called the ***logistic*** function, to model $P(y = 1)$:

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

logistic curve

linear component of some transformed scale

Logistic Regression

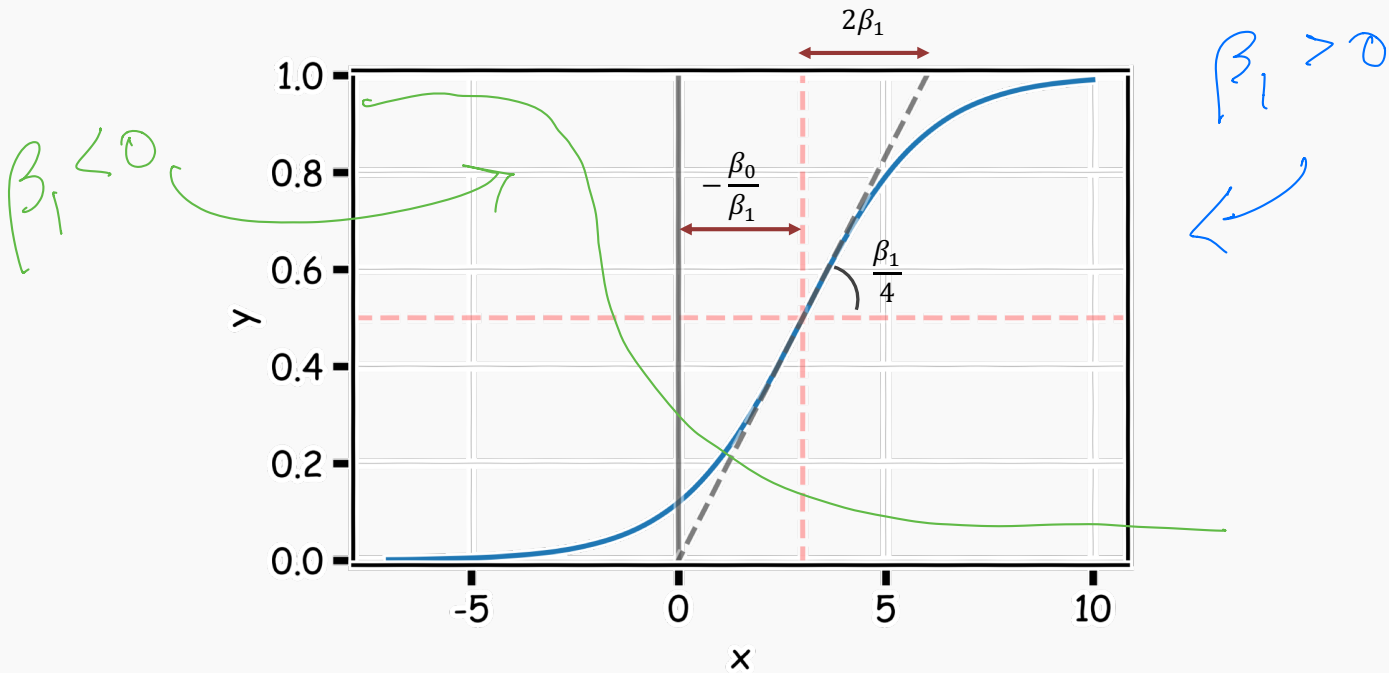
As a result the model will predict $P(y = 1)$ with an S-shaped curve, which is the general shape of the logistic function.

β_0 shifts the curve right or left by c $= -\frac{\beta_0}{\beta_1}$.

β_1 controls how steep the S-shaped curve is. Distance from ½ to almost 1 or ½ to almost 0 to ½ is $\frac{2}{\beta_1}$

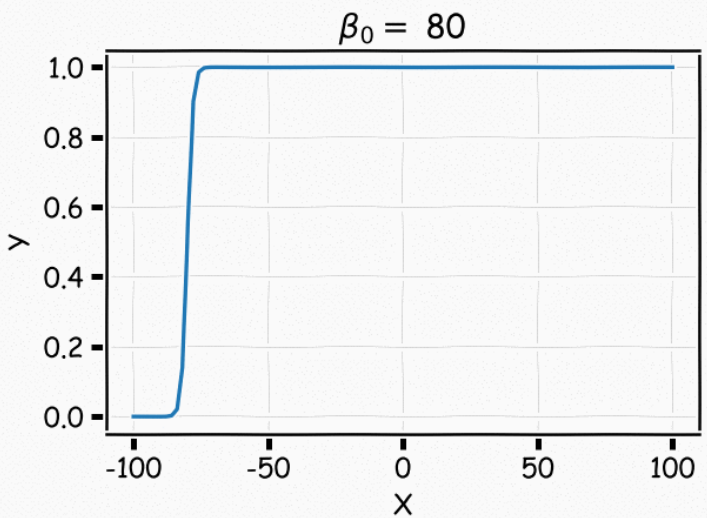
Note: if β_1 is positive, then the predicted $P(y = 1)$ goes from zero for small values of X to one for large values of X and if β_1 is negative, then the $P(y = 1)$ has opposite association.

Logistic Regression



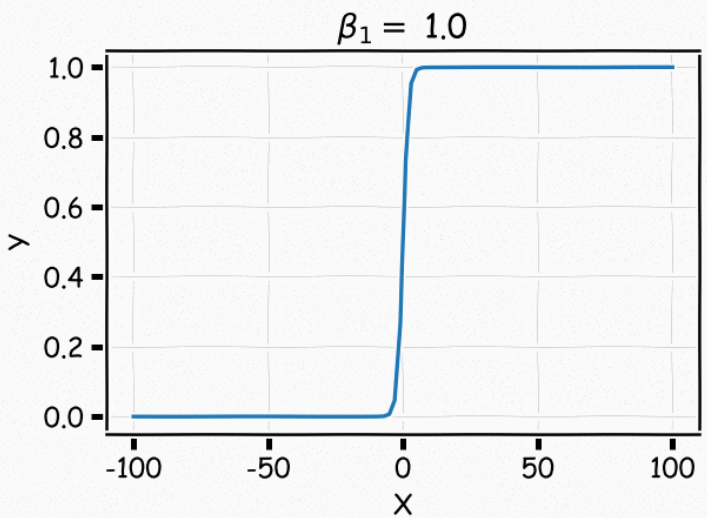
Logistic Regression

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$



Logistic Regression

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$



Logistic Regression

With a little bit of algebraic work, the logistic model can be rewritten as:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X.$$

The value inside the natural log function $\frac{P(Y=1)}{1-P(Y=1)}$, is called the ***odds***, thus logistic regression is said to model the ***log-odds*** with a linear function of the predictors or features, X. This gives us the natural interpretation of the estimates similar to linear regression: a one unit change in X is associated with a β_1 change in the log-odds of $Y = 1$; or better yet, a one unit change in X is associated with an e^{β_1} change in the odds that $Y = 1$.

Estimating the Simple Logistic Model

Estimation in Logistic Regression

Unlike in linear regression where there exists a closed-form solution to finding the estimates, $\hat{\beta}_j$'s, for the true parameters, logistic regression estimates cannot be calculated through simple matrix multiplication.

MSE: minimized to get $\hat{\beta}_0, \hat{\beta}_1$

Questions:

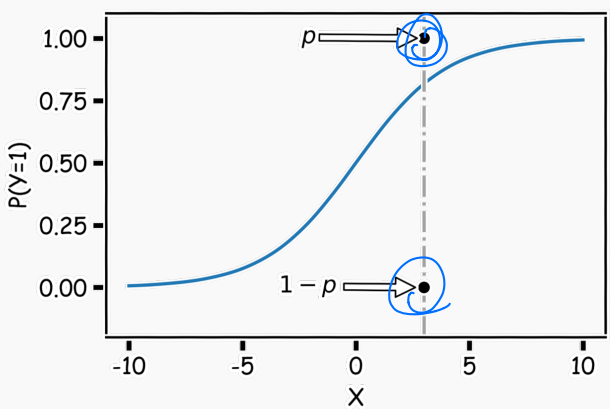
- In linear regression what loss function was used to determine the parameter estimates?
- What was the probabilistic perspective on linear regression?

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

- Logistic Regression also has a likelihood based approach to estimating parameter coefficients.

Response: 0/1 $\quad Y \sim Bern\left(p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}\right)$

Estimation in Logistic Regression



Probability $Y = 1$: p

Probability $Y = 0$: $1 - p$

$$P(Y = y) = p^{y}(1 - p)^{(1-y)}$$

where:

$p = P(Y = 1|X = x)$ and therefore p depends on X.

Thus not every p is the same for each individual measurement.

Likelihood

The likelihood of a single observation for *p* given *x and y* is:

$$L(p_i|Y_i) = P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Given the observations are independent, what is the likelihood function for *p*?

$$L(p|Y) = \prod_i P(Y_i = y_i) = \prod_i p_i^{y_i}(1-p_i)^{1-y_i}$$

$$l(p|Y) = -\log L(p|Y) = -\sum_i y_i \log p_i + (1-y_i)\log(1-p_i)$$

Loss Function

$$l(p|Y) = -\sum_i \left[y_i \log \frac{1}{1+e^{-\beta X_i}} + (1-y_i) \log\left(1 - \frac{1}{1+e^{-\beta X_i}}\right) \right]_i$$

How do we minimize this?

Differentiate, equate to zero and solve for it!

But jeeze does this look messy?! It will not necessarily have a closed form solution.

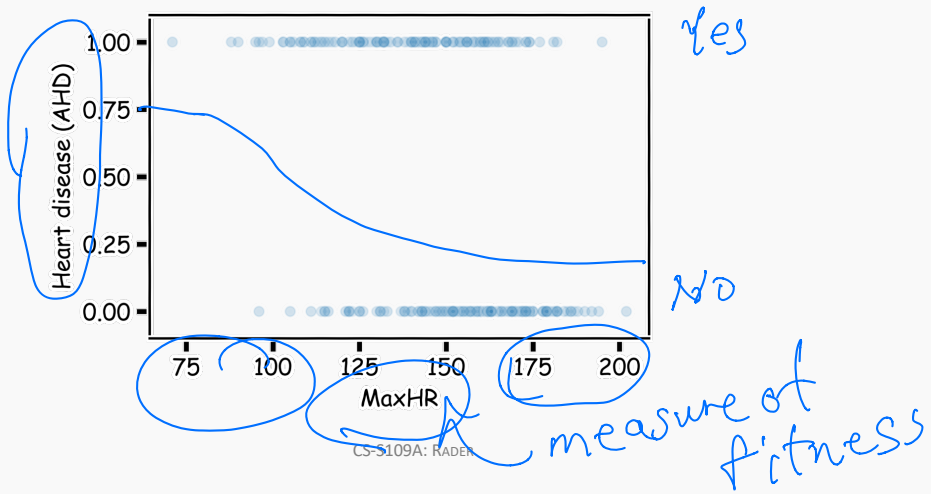
So how do we determine the parameter estimates? Through an iterative approach (we will talk about this *at length* in future lectures).

Gradient Descent

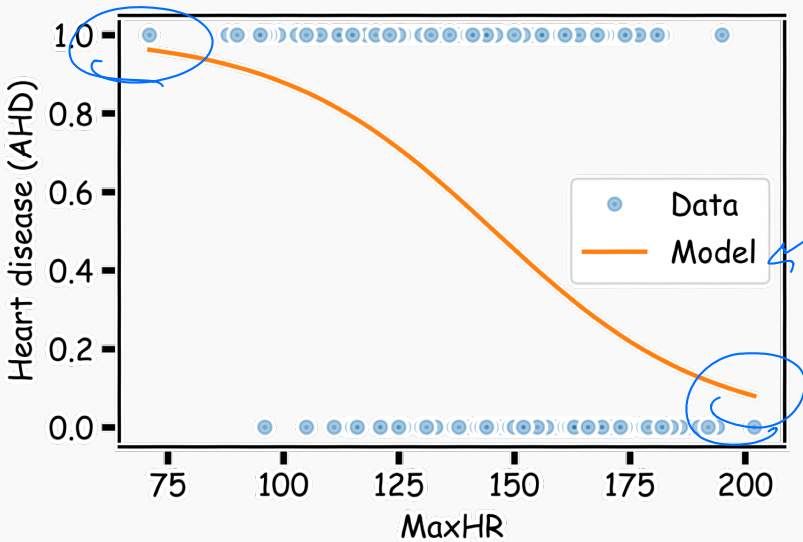
Heart Data: logistic estimation

We'd like to predict whether or not a person has a heart disease. And we'd like to make this prediction, for now, just based on the MaxHR.

How should we visualize these data?



Heart Data: logistic estimation



Predicted Probability of a heart attack

Heart Data: logistic estimation

There are various ways to fit a logistic model to this data set in Python. The most straightforward in `sklearn` is via `linear_model.LogisticRegression`.

```
from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression(C=100000, fit_intercept=True)
logreg.fit(data_x.values.reshape(-1,1), data_y);


print('Estimated beta1: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1:
 [[-0.04326016]]
Estimated beta0:
 [ 6.30193148]
```

fit to data

slope

intercept

Logistic Regression is regularized by default.

"C" = $1/\lambda$ or $1/\alpha$

Big C: unregularized.

Heart Data: logistic estimation

Answer some questions:

- Write down the logistic regression model.
- Interpret $\hat{\beta}_1$.
- Estimate the probability of heart disease for someone (like Pavlos) with MaxHR≈200?
- If we were to use this model purely for classification, how would we do so? See any issues?

$$\log\left(\frac{\hat{P}(Y=1)}{\hat{P}(Y=0)}\right) = 6.30 - 0.043(X)$$

MaxHR

the change in log-odds for a 1-unit change in X

$$\log(\widehat{\text{odds}}) = 6.30 - 0.043(200) \Rightarrow \hat{P}(Y=1) = \frac{1}{1+e^{-(6.3+0.043\times 200)}}$$

$e^{\hat{\beta}_1}$ = estimated odds ratio of "success" for a 1-unit change in X.

turn into pure classification by $\hat{P}(Y=1) > 0.5$

classification boundary condition

Categorical Predictors

Just like in linear regression, when the predictor, X, is binary, the interpretation of the model simplifies (and there is a quick closed form solution).

In this case, what are the interpretations of $\hat{\beta}_0$ and $\hat{\beta}_1$?

For the heart data, let X be the indicator that the individual is a male or female. What is the interpretation of the coefficient estimates in this case?

The observed percentage of HD for women is 26% while it is 55% for men.

Calculate the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ if the indicator for HD was predicted from the gender indicator.

For women $(X=0)$

$\log(\text{odds}) = \beta_0$

$\log\left(\frac{0.26}{0.74}\right) = \hat{\beta}_0$

For men $(X=1)$

$\log(\text{odds}) = \beta_0 + \beta_1$

$\log\left(\frac{0.55}{0.45}\right) = \hat{\beta}_0 + \hat{\beta}_1$

Statistical Inference in Logistic Regression

The **uncertainty of the estimates** $\hat{\beta}_0$ and $\hat{\beta}_1$ can be quantified and used to calculate both confidence intervals and hypothesis tests.

The estimate for the standard errors of these estimates, likelihood-based, is based on a quantity called Fisher's Information (beyond the scope of this class), which is related to the curvature of the log-likelihood function.

Due to the nature of the underlying Bernoulli distribution, if you estimate the underlying proportion p_i, you get the variance for free! Because of this, the inferences will be based on the normal approximation (and not t-distribution based). Use statsmodels to perform inferences!

Of course, you could always **bootstrap** the results to perform these inferences as well.

Classification using the Logistic Model

If $\hat{p}(Y=1) \geq 0.5 \Rightarrow$ classify to $Y=1$

$\hat{p}(Y=1) < 0.5 \Rightarrow$ classify to $Y=0$

Using Logistic Regression for Classification

How can we use a logistic regression model to perform classification?

That is, how can we predict when $Y = 1$ vs. when $Y = 0$?

We mentioned before, we can classify all observations for which $\hat{P}(Y = 1) \geq 0.5$ to be in the group associated with $Y = 1$ and then classify all observations for which $\hat{P}(Y = 0) < 0.5$ to be in the group associated with $Y = 0$.

Using such an approach is called the standard **Bayes classifier**.

the straightforward approach

The Bayes classifier takes the approach that assigns each observation to the most likely class, given its predictor values.

Using Logistic Regression for Classification

When will this Bayes classifier be a good one? When will it be a poor one?

The Bayes classifier is the one that minimizes the overall classification error rate. That is, it minimizes:

$$\frac{1}{n}\sum_{i}^{n} I\left(y_i \neq \hat{y}_i\right)$$

misclassification error rate

Is this a good Loss function to minimize? Why or why not?

The Bayes classifier may be a poor indicator within a group. Think about the Heart Data scatter plot...

discrete. not a closed form or unique solution

Using Logistic Regression for Classification

This has potential to be a good classifier if the predicted probabilities are on both sides of 0 and 1.

← its bad if classes are severely imbalanced

How do we extend this classifier if Y has more than two categories?

↑ $Y = \{$ CS, stat, other $\}$ most likely category

Multiple Logistic Regression

Multiple Logistic Regression

It is simple to illustrate examples in logistic regression when there is just one predictors variable.

But the approach 'easily' generalizes to the situation where there are multiple predictors.

A lot of the same details as linear regression apply to logistic regression. Interactions can be considered. Multicollinearity is a concern. So is overfitting.

Etc...

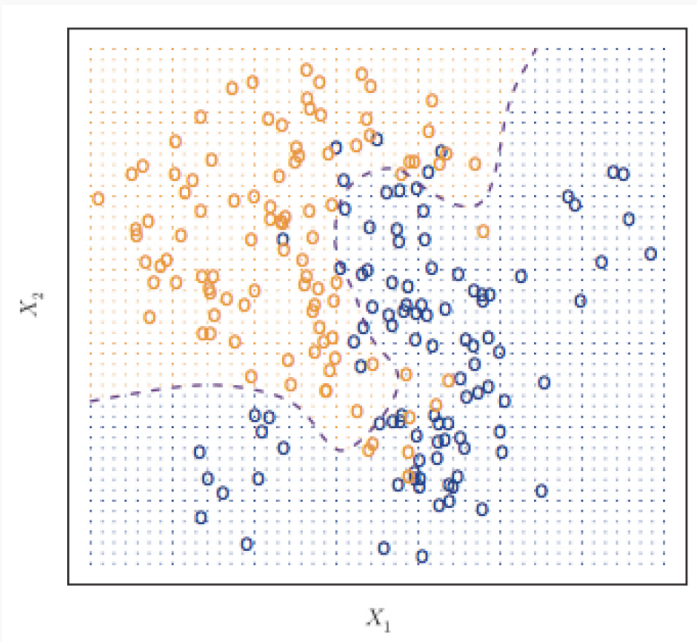
So how do we correct for such problems?

Regularization and checking though train, test, and cross-validation!

We will get into the details of this, along with other extensions of logistic regression, in the next lecture.

Classifier with two predictors

How can we estimate a classifier, based on logistic regression, for the following plot?



Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model. What was the model statement (in terms of linear predictors)?

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

Multiple logistic regression is a generalization to multiple predictors. More specifically we can define a multiple logistic regression model to predict $P(Y=1)$ as such:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

holding other predictors constant

Fitting Multiple Logistic Regression

The estimation procedure is identical to that as before for simple logistic regression:

- a likelihood approach is taken, and the function is maximized across all parameters $\beta_0, \beta_1, \ldots, \beta_p$ using an iterative method like Newton-Raphson or Gradient Descent.

The actual fitting of a Multiple Logistic Regression is easy using software (of course there's a python package for that) as the iterative maximization of the likelihood has already been hard coded.

In the `sklearn.linear_model` package, you just have to create your multidimensional design matrix X to be used as predictors in the `LogisticRegression` function.

Interpretation of Multiple Logistic Regression

Interpreting the coefficients in a multiple logistic regression is similar to that of linear regression.

Key: since there are other predictors in the model, the coefficient $\hat{\beta}_j$ is the association between the j^{th} predictor and the response (on log odds scale). But do we have to say? Controlling for the other predictors in the model.

We are trying to attribute the partial effects of each model controlling for the others (aka, controlling for possible *confounders*}.

↑ worry if multicollinearity is present.

Interpreting Multiple Logistic Regression: an Example

Let's get back to the Heart Data. We are attempting to predict whether someone has HD based on MaxHR and whether the person is female or male. The simultaneous effect of these two predictors can be brought into one model.

Recall from earlier we had the following estimated models:

$$\log\left(\frac{\widehat{P(Y=1)}}{1-\widehat{P(Y=1)}}\right) = 6.30 - 0.043 \cdot X_{MaxHR}$$

$$\log\left(\frac{\widehat{P(Y=1)}}{1-\widehat{P(Y=1)}}\right) = -1.06 + 1.27 \cdot X_{gender}$$

$\hat{\beta}_0 < 0$: women's P(heart attack) < 0.5

$\hat{\beta}_1 > 0$: men's probabilities are higher than women

Interpreting Multiple Logistic Regression: an Example

The results for the multiple logistic regression model are:

```
data_x = df_heart[['MaxHR','Sex']]
data_y = df_heart['AHD']

logreg = LogisticRegression(C=100000, fit_intercept=True)
logreg.fit(data_x, data_y);


print('Estimated beta1: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1:
 [[-0.04496354  1.40079047]]
Estimated beta0:
 [ 5.58662464]
```

$\hat{\beta}_{maxHR} = -0.045$

$\hat{\beta}_{gender} = 1.400$

Some questions

1. Estimate the odds ratio of HD comparing men to women using this model.
2. Is there any evidence of multicollinearity in this model?
3. Is there any confounding in this problem?

Interactions in Multiple Logistic Regression

Just like in linear regression, interaction terms can be considered in logistic regression. An **interaction terms** is incorporated into the model the same way, and the interpretation is very similar (on the log-odds scale of the response of course).

Write down the model for the Heart data for the 2 predictors plus the interactions term.

Interpreting Multiple Logistic Regression: an Example

The results for the multiple logistic regression model are:

Slide Type -

```
df_heart['Interaction'] = df_heart.MaxHR * df_heart.Sex

data_x = df_heart[['MaxHR','Sex', 'Interaction']]
data_y = df_heart['AHD']

logreg = LogisticRegression(C=100000, fit_intercept=True)
logreg.fit(data_x, data_y);


print('Estimated beta1, beta2, beta3: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1, beta2, beta3:
 [[-0.02645985  5.38749287 -0.02689767]]
Estimated beta0:
 [ 2.88218441]
```

measuring the differences in maxHR association with "AHD" comparing men to women.

Some questions

1. Write down the complete model. Break this down into the model to predict log-odds of heart disease (HD) based on MaxHR for women and the same model for men. How is this different from the previous model (without interaction)?

2. Interpret the results of this model. What does the coefficient for the interaction term represent?

3. Estimate the odds ratio of HD comparing men to women using this model [trick question].

4. Is there any evidence of multicollinearity in this model?

5. Is there any confounding in this problem?

Extending the Logistic Model

Assumptions ?

Model Diagnostics in Logistic Regression

In linear regression, when is the model appropriate (aka, what are the assumptions)?

In logistic regression, when is the model appropriate?

We don't have to worry about the distribution of the residuals (we get that for free).

What we do have to worry about is how Y 'links' to X in its relationship. More specifically, we assume the 'S'-shaped (aka, sigmoidal) curve follows the logistic function. How could we check this?

Alternatives to logistic regression

Why was the logistic function chosen to model how a binary response variable can be predicted from a quantitative predictor?

Because it it takes as inputs values in $(0,1)$ and outputs values $(-\infty, \infty)$ so that the estimation of β is unbounded.

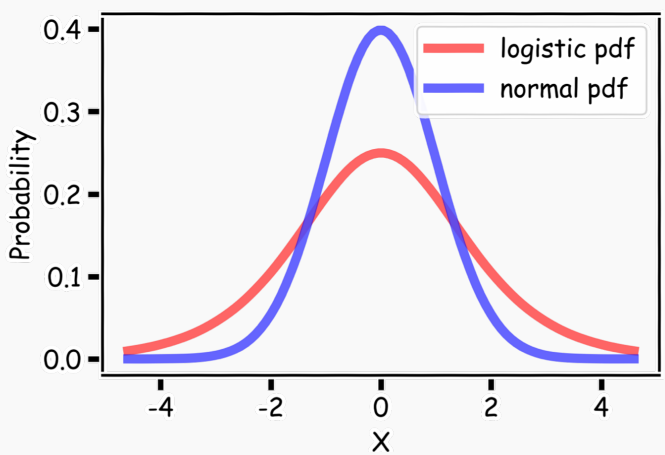
This is not the only function that does this. Any suggestions?

Any *inverse CDF* function for an unbounded continuous distribution can work as the `link' between the observed values for $Y$ and how it relates `linearly' to the predictors.

So what are possible other choices? What differences do they have? Why is logistic regression preferred?

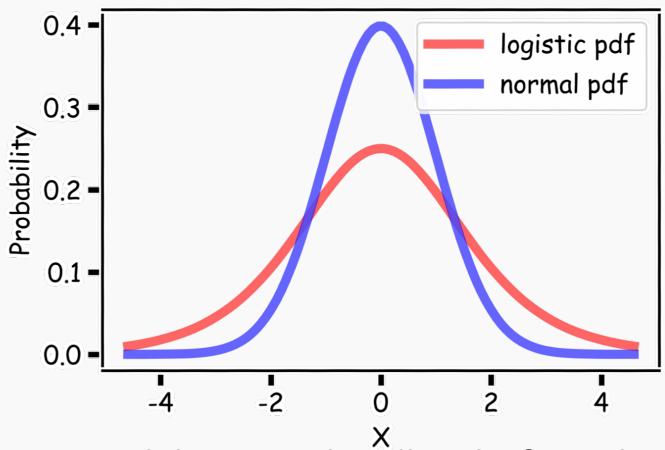
Logistic vs Normal pdf

The choice of link function determines the shape of the S' shape. Let's compare the pdf's for the Logistic and Normal distributions (called a 'probit' model, econometricians love these):



So what?

Logistic vs Normal pdf



Choosing a distribution with longer tails will make for a shape that asymptotes more slowly (likely a good thing for model fitting).

Classification Boundaries

Classification boundaries

Recall that we could attempt to purely classify each observation based on whether the estimated $P(Y = 1)$ from the model was greater than 0.5.

When dealing with 'well-separated' data, logistic regression can work well in performing classification.

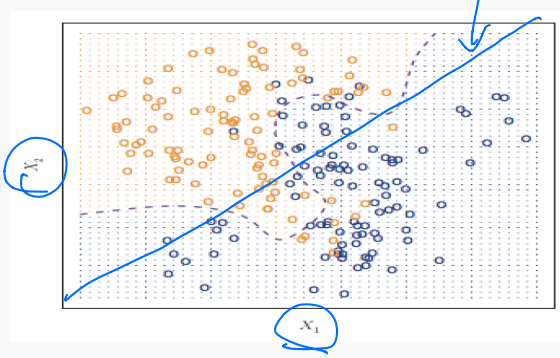
We saw a 2-D plot last time which had two predictors, X_1, X_2 and depicted the classes as different colors. A similar one is shown on the next slide.

$\log(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

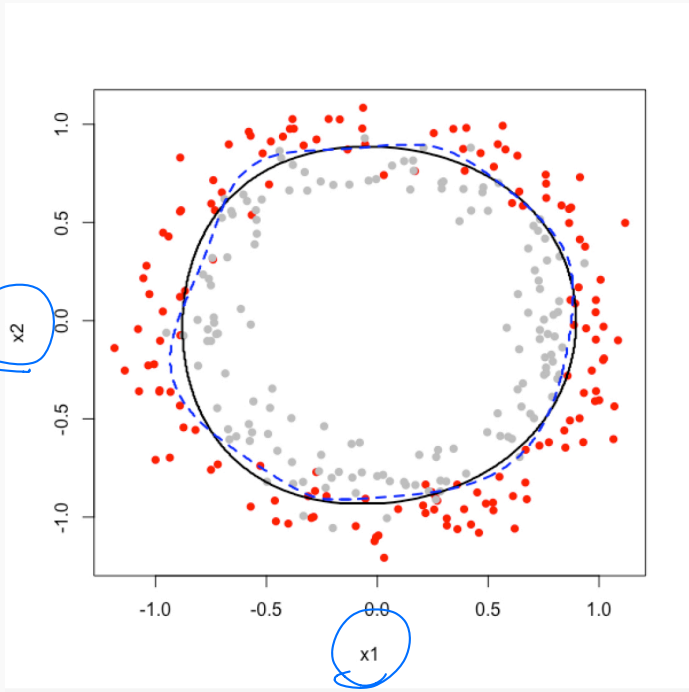
linear boundary

$Y=1$ (orange)

$Y=0$ (blue)



2D Classification in Logistic Regression: an Example



red $\Rightarrow Y=1$

gray $\Rightarrow Y=0$

polynomial of order 2

$\text{log-odds} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \beta_5 (X_1 \cdot X_2)$

2D Classification in Logistic Regression: an Example

Would a logistic regression model perform well in classifying the observations in this example?

What would be a good logistic regression model to classify these points?

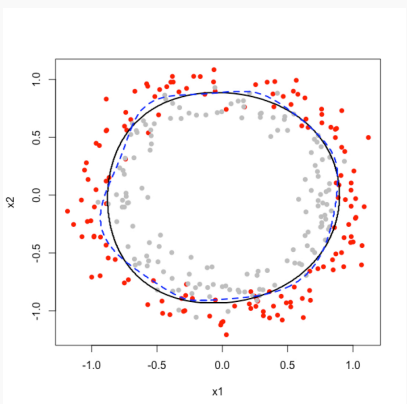
Based on these predictors, two separate logistic regression model were considered that were based on different ordered polynomials of X_1, X_2 and their interactions. The 'circles' represent the boundary for classification.

How can the classification boundary be calculated for a logistic regression?

2D Classification in Logistic Regression: an Example

In the previous plot, which classification boundary performs better? How can you tell? How would you make this determination in an actual data example?

We could determine the misclassification rates in left out validation or test set(s)

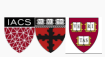


Polynomial logistic regression of order 2

Lecture Outline

- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
  - Estimating the Simple Logistic Model
  - Classification using the Logistic Model
  - Multiple Logistic Regression
  - Extending the Logistic Model
- Classification Boundaries
- **Regularization in Logistic Regression**
- Multinomial Logistic Regression
- Bayes Theorem and Misclassification Rates
- ROC Curves

prevent overfitting
control for multicolinearity



Review: Regularization in Linear Regression

Based on the Likelihood framework, a loss function can be determined based on the likelihood function.

We saw in linear regression that maximizing the log-likelihood is equivalent to minimizing the sum of squares error:

$$\arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}))^2$$

Review: Regularization in Linear Regression

And a regularization approach was to add a penalty factor to this equation. Which for Ridge Regression becomes:

$$\arg\min \left[\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{n} \beta_j x_{ji} \right) \right)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \right]$$

penalty as a function of β magnitudes

Note: this penalty *shrinks* the estimates towards zero, and had the analogue of using a Normal prior centered at zero in the Bayesian paradigm.

Loss function in Logistic Regression

penalize this loss function

A similar approach can be used in logistic regression. Here, maximizing the log-likelihood is equivalent to minimizing the following loss function:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \left[-\sum_{i=1}^{n} \left(y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \right) \right]$$

where $p_i = \frac{1}{1-e^{-(\beta_0+\beta_1 X_{1,i}+\cdots+\beta_p x_{p,i})}}$

Why is this a good loss function to minimize? Where does this come from?

The log-likelihood for independent $Y_i \sim \text{Bern}(p_i)$.

Regularization in Logistic Regression

A penalty factor can then be added to this loss function and results in a new loss function that penalizes large values of the parameters:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \left[-\sum_{i=1}^{n} \left(y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

The result is just like in linear regression: shrink the parameter estimates towards zero.

In practice, the intercept is usually not part of the penalty factor.

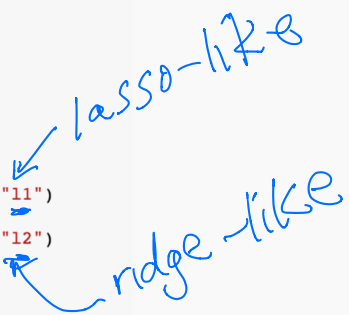
Note: the sklearn package uses a different tuning parameter: instead of λ they use a constant that is essentially $C = \frac{1}{\lambda}$.

sklearn's implementation

Regularization in Logistic Regression: an Example

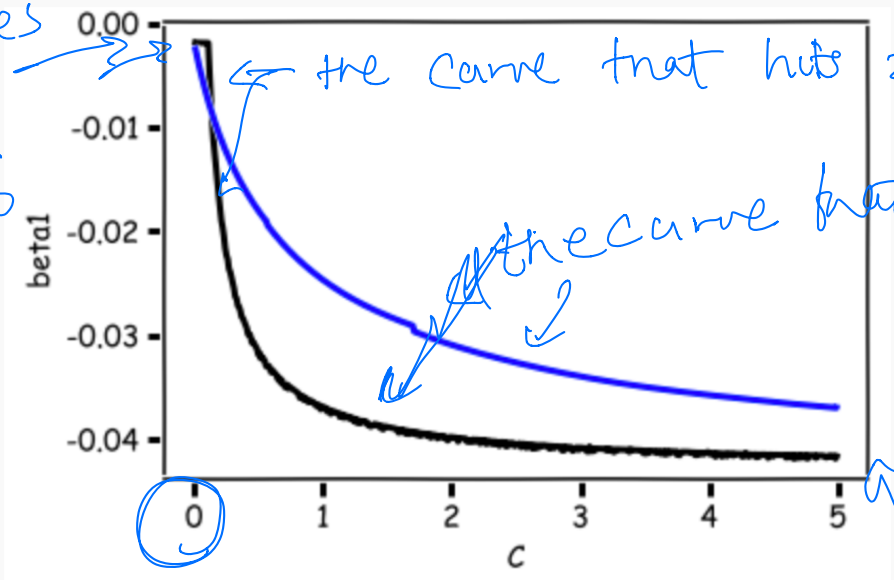
Let's see how this plays out in an example in logistic regression.

```
beta1_l1 = []
beta1_l2 = []
Cs = []
data_x = df_heart[['MaxHR']]
data_y = df_heart['AHD']

for i in range(1, 500):
    C = i/100
    logitm_l1 = sk.LogisticRegression(C = C, penalty = "l1")
    logitm_l1.fit (data_x, data_y)
    logitm_l2 = sk.LogisticRegression(C = C, penalty = "l2")
    logitm_l2.fit (data_x, data_y)
    beta1_l1.append(logitm_l1.coef_[0])
    beta1_l2.append(logitm_l2.coef_[0])
    Cs.append(C)

plt.plot(Cs, beta1_l1,  color='black', lw=3)
plt.plot(Cs, beta1_l2,  color='blue', lw=3)
plt.xlabel ("C")
plt.ylabel("beta1")
plt.show()
```



Regularization in Logistic Regression: an Example

Let's see how this plays out in an example in logistic regression.



estimates are shrunk to zero

the curve that hits zero is L1 (Lasso)

the curve that asymptotes towards zero is Ridge

C is big, we approach standard logistic regression estimates

Regularization in Logistic Regression: tuning λ

Just like in linear regression, the shrinkage factor must be chosen. How should we go about doing this?

Through building multiple training and test sets (through k-fold or random subsets), we can select the best shrinkage factor to mimic out-of-sample prediction.

How could we measure how well each model fits the test set?

We could measure this based on some loss function (more coming later)!

Lecture Outline

- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
  - Estimating the Simple Logistic Model
  - Classification using the Logistic Model
  - Multiple Logistic Regression
  - Extending the Logistic Model
- Classification Boundaries
- Regularization in Logistic Regression
- **Multinomial Logistic Regression** ← when Y has 3+ groups
- Bayes Theorem and Misclassification Rates
- ROC Curves

Logistic Regression for predicting more than 2 Classes

There are several extensions to standard logistic regression when the response variable Y has more than 2 categories. The two most common are:

1. ordinal logistic regression
2. multinomial logistic regression.

Ordinal logistic regression is used when the categories have a specific hierarchy (like class year: Freshman, Sophomore, Junior, Senior; or a 7-point rating scale from strongly disagree to strongly agree).

Multinomial logistic regression is used when the categories have no inherent order (like eye color: blue, green, brown, hazel, et...).

Multinomial Logistic Regression

There are two common approaches to estimating a nominal (not-ordinal) categorical variable that has more than 2 classes. The first approach sets one of the categories in the response variable as the *reference* group, and then fits separate logistic regression models to predict the other cases based off of the reference group. For example we could attempt to predict a student's concentration:

$$y = \begin{cases} 1 & if \text{ Computer Science (CS)} \\ 2 & if \text{ Statistics} \\ 3 & \text{otherwise} \end{cases}$$

from predictors x_1 number of psets per week and x_2 how much time spent in Lamont Library.

Multinomial Logistic Regression (cont.)

We could select the $y = 3$ case as the reference group (other concentration), and then fit two separate models: a model to predict $y = 1$ (CS) from $y = 3$ (others) and a separate model to predict $y = 2$ (Stat) from $y = 3$ (others).

Ignoring interactions, how many parameters would need to be estimated?

How could these models be used to estimate the probability of an individual falling in each concentration?

doubles that in a standard logistic regression models

↑ solving for a system of 3 equations in this case.

1. $\sum p = 1$
2. $P(Y=3)$ compares for $P(Y=1)$
3. " " " $P(Y=2)$

One vs. Rest (ovr) Logistic Regression

The default multiclass logistic regression model is called the 'One vs. Rest' approach, which is our second method.

If there are 3 classes, then 3 separate logistic regressions are fit, where the probability of each category is predicted over the rest of the categories combined. So for the concentration example, 3 models would be fit:

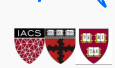
- a first model would be fit to predict CS from (Stat and Others) combined.
- a second model would be fit to predict Stat from (CS and Others) combined.
- a third model would be fit to predict Others from (CS and Stat) combined.

An example to predict play call from the NFL data follows...

OVR Logistic Regression in Python

```
#read the NFL play-by-play data
nfldata = pd.read_csv("NFLplaybyplay-2015.csv")

# shuffle the data
nfldata = nfldata.reindex(np.random.permutation(nfldata.index))

# For simplicity, we will select only 500 points form the dataset.
N = 500

X = nfldata[["YardLine"]]

nfldata["PlayType"]=nfldata["IsPass"]+2*nfldata["IsRush"]

logitm = sk.LogisticRegression(C = 10000000)
logitm.fit (X, nfldata["PlayType"])

# The coefficients
print('Estimated beta1: \n', logitm.coef_)
print('Estimated beta0: \n', logitm.intercept_)
```

```
Estimated beta1:
 [[-0.01460736]
 [ 0.00635893]
 [ 0.00652455]]
Estimated beta0:
 [-0.26422696 -0.61186328 -1.20051275]
```



Classification for more than 2 Categories

When there are more than 2 categories in the response variable, then there is no guarantee that $P(Y = k) \geq 0.5$ for any one category. So any classifier based on logistic regression will instead have to select the group with the largest estimated probability.

(our approach makes this easy)

The classification boundaries are then much more difficult to determine. We will not get into the algorithm for drawing these in this class.

Classification for more than 2 Categories

When there are more than 2 categories in the response variable, then there is no guarantee that $P(Y = k) \geq 0.5$ for any one category. So any classifier based on logistic regression will instead have to select the group with the largest estimated probability.

The classification boundaries are then much more difficult to determine. We will not get into the algorithm for drawing these in this class.

Softmax

So how do we convert a set of probability estimates from separate models to one set of probability estimates?

The ***softmax*** function is used. That is, the weights are just normalized for each predicted probability. AKA, predict the 3 class probabilities from each model of the 3 models, and just rescale so they add up to 1.

Mathematically that is:

$$P(y = k|\vec{x}) = \frac{e^{\vec{x}^T \hat{\vec{\beta}}_k}}{\sum_{j=1}^{K} e^{\vec{x}^T \hat{\vec{\beta}}_j}}$$

where $\vec{x}$ is the vector of covariates for that observation and $\hat{\vec{\beta}}_k$ are the associated logistic regression coefficient estimates.

Classification for more than 2 Categories in sklearn

```
X=np.arange(100)
print(logitm.predict_proba(X.reshape(-1,1))[0:10,:])
print(logitm.predict(X.reshape(-1,1)))
```

```
[[ 0.42692074  0.34563977  0.22743948]
 [ 0.42380137  0.34739801  0.22880062]
 [ 0.42067735  0.34915746  0.23016519]
 [ 0.41754902  0.35091792  0.23153306]
 [ 0.41441671  0.35267918  0.23290411]
 [ 0.41128078  0.35444103  0.23427819]
 [ 0.40814155  0.35620326  0.23565519]
 [ 0.40499938  0.35796565  0.23703497]
 [ 0.40185462  0.35972799  0.23841739]
 [ 0.39870763  0.36149006  0.23980231]]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
```

group 2 never shows up as the highest probability

Lecture Outline

- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
  - Estimating the Simple Logistic Model
  - Classification using the Logistic Model
  - Multiple Logistic Regression
  - Extending the Logistic Model
- Classification Boundaries
- Regularization in Logistic Regression
- Multinomial Logistic Regression
- **Bayes Theorem and Misclassification Rates**
- ROC Curves (AUC)

} evaluating classification models

Bayes' Theorem

We defined conditional probability as:

$$P(B|A) = P(B \text{ and } A)/P(A)$$

And using the fact that $P(B \text{ and } A) = P(A|B)P(B)$ we get the simplest form of Bayes' Theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Another version of Bayes' Theorem is found by substituting in the Law of Total Probability (LOTP) into the denominator:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}$$

Where have we seen Bayes' Theorem before? Why do we care?

↑ hopefully in intro statistics

Diagnostic Testing

In the diagnostic testing paradigm, one cares about whether the results of a test (like a classification test) matches truth (the true class that observation belongs to). The simplest version of this is trying to detect disease ($D+$ vs. $D-$) based on a diagnostic test ($T+$ vs. $T-$).

Medical examples of this include various screening tests: breast cancer screening through (i) self-examination and (ii) mammography, prostate cancer screening through (iii) PSA tests, and Colo-rectal cancer through (iv) colonoscopies.

These tests are a little controversial because of poor predictive probability of the tests.

Diagnostic Testing (cont.)

Bayes' theorem can be rewritten for diagnostic tests:

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

These probability quantities can then be defined as:

- *Sensitivity*: $P(T+|D+)$
- *Specificity*: $P(T-|D-)$
- *Prevalence*: $P(D+)$
- *Positive Predictive Value*: $P(D+|T+)$
- *Negative Predictive Value*: $P(D-|T-)$

How do positive and negative predictive values relate? Be careful...

Diagnostic Testing

We mentioned that these tests are a little controversial because of their poor predictive probability. When will these tests have poor positive predictive probability?

When the disease is not very prevalent, then the number of 'false positives' will overwhelm the number of true positive. For example, PSA screening for prostate cancer has sensitivity of about 90% and specificity of about 97% for some age groups (men in their fifties), but prevalence is about 0.1%. ← if prevalence is low, then PPV is not high

What is positive predictive probability for this diagnostic test?

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Because classification can be thought of as a diagnostic test. Let $Y_i = k$ be the event that observation *i* truly belongs to category *k*, and let $\hat{Y}_i = k$ the event that we correctly predict it to be in class *k*. Then Bayes' rule states that our *Positive Predictive Value* for classification is:

$$P(Y_i = k|\hat{Y}_i = k) = \frac{P(\hat{Y}_i = k|Y_i = k)P(Y_i = k)}{P(\hat{Y}_i = k|Y_i = k)P(Y_i = k) + P(\hat{Y}_i = k|Y_i \neq k)P(Y_i \neq k)}$$

Thus the probability of a predicted outcome truly being in a specific group depends on what? The proportion of observations in that class!

Error in Classification

There are 2 major types of error in classification problems based on a binary outcome. They are:

False positives: incorrectly predicting $\hat{Y} = 1$ when it truly is in $Y = 0$.

False negative: incorrectly predicting $\hat{Y} = 0$ when it truly is in $Y = 1$.

The results of a classification algorithm are often summarized in two ways: a confusion table, sometimes called a contingency table, or a 2x2 table (more generally (*k* x *k*) table) and a receiver operating characteristics (ROC) curve.

Confusion table

When a classification algorithm (like logistic regression) is used, the results can be summarize in a (k x k) table as such:

	Predicted not Republican ($\hat{Y} = 0$)	Predicted Republican ($\hat{Y} = 1$)
Truly not Republican ($Y = 0$)	288	487
Truly Republican ($Y = 0$)	311	221

confusion matrix

The table above was a classification based on a logistic regression model to predict political party (Dem. vs. Rep.) based on 3 predictors: X_1 = whether respondent believes abortion is legal, X_2 = income (logged) and X_3 = years of education.

What are the false positive and false negative rates for this classifier?

$$P(Y=0 \mid \hat{Y}=1) = \frac{P(Y=0 \text{ and } \hat{Y}=1)}{P(\hat{Y}=1)} = \frac{487}{487+221} = \frac{487}{708} \approx 70\%$$

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table. How?

The choice of the Bayes' classifier level will modify the characteristics of this table.

If we thought is was more important to predict republicans correctly (lower false positive rate), what could we do for our Bayes' classifier level?

We could classify instead based on:

$$\hat{P}(Y = 1) < \pi$$

instead of 0.5

and we could choose π to be some level other than 0.5. Let's see what the table looks like if π were 0.48 or 0.72 instead (why such strange numbers?).

↑ by changing the classification threshold, we can alter FPR, FNR

Other Confusion tables

Based on π = 0.48:

	Predicted not Republican ($\hat{Y} = 0$)	Predicted Republican ($\hat{Y} = 1$)
Truly not Republican ($Y = 0$)	143	632
Truly Republican ($Y = 0$)	138	394

What has improved? What has worsened?

Based on π = 0.72:

	Predicted not Republican ($\hat{Y} = 0$)	Predicted Republican ($\hat{Y} = 1$)
Truly not Republican ($Y = 0$)	539	236
Truly Republican ($Y = 0$)	455	77

Which should we choose? Why?

have very different rates of false positives and false negatives

Lecture Outline

- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
  - Estimating the Simple Logistic Model
  - Classification using the Logistic Model
  - Multiple Logistic Regression
  - Extending the Logistic Model
- Classification Boundaries
- Regularization in Logistic Regression
- Multinomial Logistic Regression
- Bayes Theorem and Misclassification Rates
- **ROC Curves**

Illustrates the trade-off between TPR and FPR for any classification algorithm

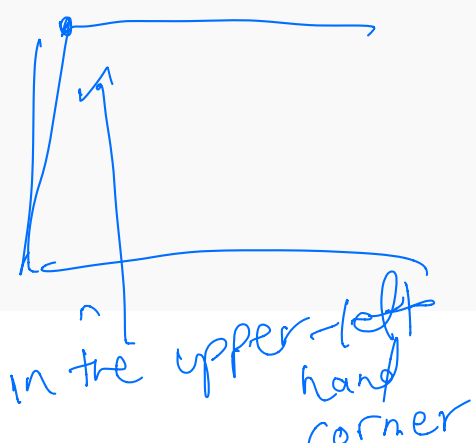
ROC Curves

The Radio Operator Characteristics (ROC) curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

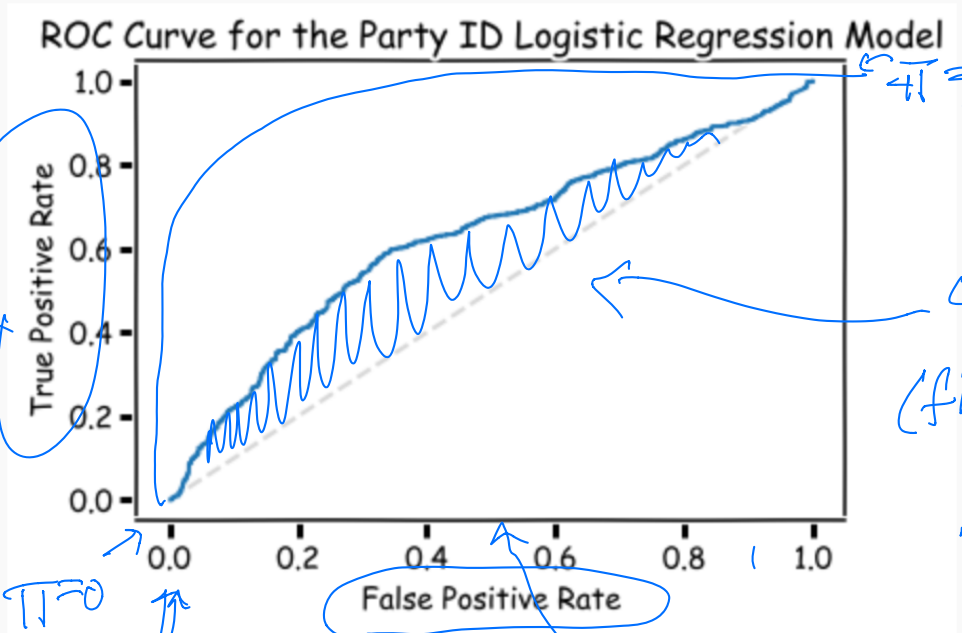
The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

See next slide for an example.



ROC Curve Example



AUC: measures how strong the model is in comparison to the baseline of random assignment.

ROC Curve for measuring classifier performance

The overall performance of a classifier, calculated over all possible thresholds, is given by the area under the ROC curve ('AUC').

An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

What is the worst case scenario for AUC? What is the best case? What is AUC if we independently just flip a coin to perform classification?

This AUC then can be use to compare various approaches to classification: Logistic regression, LDA (to come), *k*-NN, etc...