Logistic Regression
Why use logistic regression?

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1 (did vote)
The Linear Probability Model

In the OLS regression:

\[ Y = \gamma + \beta X + e ; \text{ where } Y = (0, 1) \]

- The error terms are heteroskedastic
- \( e \) is not normally distributed because \( Y \) takes on only two values
- The predicted probabilities can be greater than 1 or less than 0
Q: EVAC

Did you evacuate your home to go someplace safer before Hurricane Dennis (Floyd) hit?

1 YES
2 NO
3 DON'T KNOW
4 REFUSED
## The Data

<table>
<thead>
<tr>
<th>EVAC</th>
<th>PETS</th>
<th>MOBLHOME</th>
<th>TENURE</th>
<th>EDUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>26</td>
<td>12</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>12</td>
</tr>
</tbody>
</table>
## OLS Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>0.190</td>
<td>2.121</td>
</tr>
<tr>
<td>PETS</td>
<td>-0.137</td>
<td>-5.296</td>
</tr>
<tr>
<td>MOBLHOME</td>
<td>0.337</td>
<td>8.963</td>
</tr>
<tr>
<td>TENURE</td>
<td>-0.003</td>
<td>-2.973</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.003</td>
<td>0.424</td>
</tr>
<tr>
<td>FLOYD</td>
<td>0.198</td>
<td>8.147</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.145</td>
<td></td>
</tr>
<tr>
<td>F-stat</td>
<td>36.010</td>
<td></td>
</tr>
</tbody>
</table>
### Problems:

**Predicted Values outside the 0,1 range**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstandardized Predicted Value</td>
<td>1070</td>
<td>-.0849</td>
<td>.7602</td>
<td>.242990</td>
<td>.1632534</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>1070</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Heteroskedasticity

Dependent Variable: LNESQ
B  t-stat
(Constant) -2.34  -15.99
LNTNSQ -0.20  -6.19

Park Test
The Logistic Regression Model

The "logit" model solves these problems:

\[ \ln\left(\frac{p}{1-p}\right) = \alpha + \beta X + e \]

- \( p \) is the probability that the event Y occurs, \( p(Y=1) \)
- \( \frac{p}{1-p} \) is the "odds ratio"
- \( \ln[p/(1-p)] \) is the log odds ratio, or "logit"
More:

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

\[ p = \frac{1}{1 + \exp(-\alpha - \beta X)} \]

- If you let \( \alpha + \beta X = 0 \), then \( p = 0.50 \)
- As \( \alpha + \beta X \) gets really big, \( p \) approaches 1
- As \( \alpha + \beta X \) gets really small, \( p \) approaches 0
Comparing LP and Logit Models
Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures the probability of observing the particular set of dependent variable values \( (p_1, p_2, ..., p_n) \) that occur in the sample:
  \[
  L = \text{Prob} (p_1 \ast p_2 \ast \ast \ast p_n)
  \]
- The higher the L, the higher the probability of observing the ps in the sample.
- MLE involves finding the coefficients \((\alpha, \beta)\) that makes the log of the likelihood function \((LL < 0)\) as large as possible

- Or, finds the coefficients that make \(-2\) times the log of the likelihood function \((-2LL)\) as small as possible

- The maximum likelihood estimates solve the following condition:

\[
\{Y - p(Y=1)\}X_i = 0
\]

summed over all observations, \(i = 1,\ldots, n\)
Classification

- **Learn**: \( h: X \rightarrow Y \)
  - \( X \) – features
  - \( Y \) – target classes

- Suppose you know \( P(Y | X) \) exactly, how should you classify?
  - Bayes classifier:

- Why?
  \[
y^* = h_{\text{bayes}}(x) = \arg \max_y P(Y = y | X = x)
\]
Generative vs. Discriminative Classifiers - Intuition

- Generative classifier, e.g., Naïve Bayes:
  - Assume some functional form for $P(X \mid Y), P(Y)$
  - Estimate parameters of $P(X \mid Y), P(Y)$ directly from training data
  - Use Bayes rule to calculate $P(Y \mid X=x)$
  - This is ‘generative’ model
    - Indirect computation of $P(Y \mid X)$ through Bayes rule
    - But, can generate a sample of the data, $P(X) = \sum_y P(y)P(X \mid y)$

- Discriminative classifier, e.g., Logistic Regression:
  - Assume some functional form for $P(Y \mid X)$
  - Estimate parameters of $P(Y \mid X)$ directly from training data
  - This is the ‘discriminative’ model
    - Directly learn $P(Y \mid X)$
    - But cannot sample data, because $P(X)$ is not available
The Naïve Bayes Classifier

- Given:
  - Prior $P(Y)$
  - $n$ conditionally independent features $X$ given the class $Y$
  - For each $X_i$, we have likelihood $P(X_i | Y)$

- Decision rule:
  
  $$ y^* = h_{NB}(x) = \arg \max_y P(y)P(x_1, ..., x_n | y) $$
  
  $$ = \arg \max_y P(y)\prod P(x_i | y) $$

- If assumption holds, NB is optimal classifier!
Logistic Regression

- Let $X$ be the data instance, and $Y$ be the class label: Learn $P(Y \mid X)$ directly
  - Let $W = (W_1, W_2, \ldots, W_n)$, $X=(X_1, X_2, \ldots, X_n)$, $WX$ is the dot product
  - Sigmoid function:
    \[
    P(Y = 1 \mid X) = \frac{1}{1 + e^{-wx}}
    \]
Logistic Regression

- In logistic regression, we learn the conditional distribution $P(y \mid x)$.
- Let $p_y(x; w)$ be our estimate of $P(y \mid x)$, where $w$ is a vector of adjustable parameters.
- Assume there are two classes, $y = 0$ and $y = 1$ and

$$p_1(x; w) = \frac{1}{1 + e^{-wx}} \quad p_0(x; w) = 1 - \frac{1}{1 + e^{-wx}}$$

- This is equivalent to

$$\log \frac{p_1(x; w)}{p_0(x; w)} = wx$$

- That is, the log odds of class 1 is a linear function of $x$.
- Q: How to find $w$?
Constructing a Learning Algorithm

- The conditional data likelihood is the probability of the observed $Y$ values in the training data, conditioned on their corresponding $X$ values. We choose parameters $w$ that satisfy

$$w = \arg \max \prod_l P(y^l | x^l, w)$$

- where $w = <w_0, w_1, ..., w_n>$ is the vector of parameters to be estimated, $y^l$ denotes the observed value of $Y$ in the $l$th training example, and $x^l$ denotes the observed value of $X$ in the $l$th training example.
Equivalently, we can work with the log of the conditional likelihood:

\[
\mathbf{w} = \arg \max_{\mathbf{w}} \sum_{l} \ln P(y^l \mid \mathbf{x}^l, \mathbf{w})
\]

This conditional data log likelihood, which we will denote \( l(\mathbf{w}) \) can be written as

\[
l(\mathbf{w}) = \sum_{l} y^l \ln P(y^l = 1 \mid \mathbf{x}^l, \mathbf{w}) + (1 - y^l) \ln P(y^l = 0 \mid \mathbf{x}^l, \mathbf{w})
\]

Note here we are utilizing the fact that \( Y \) can take only values 0 or 1, so only one of the two terms in the expression will be non-zero for any given \( y^l \).
Computing the Likelihood

- We can re-express the log of the conditional likelihood as:

\[
l(w) = \sum_l y^l \ln P(y^l = 1 | x^l, w) + (1 - y^l) \ln P(y^l = 0 | x^l, w)
\]

\[
= \sum_l y^l \ln \frac{P(y^l = 1 | x^l, w)}{P(y^l = 0 | x^l, w)} + \ln P(y^l = 0 | x^l, w)
\]

\[
= \sum_l y^l (w_0 + \sum_{i=1}^n w_i x^l_i) - \ln(1 + \exp(w_0 + \sum_{i=1}^n w_i x^l_i))
\]
Fitting LR by Gradient Ascent

- Unfortunately, there is no closed form solution to maximizing \( l(w) \) with respect to \( w \). Therefore, one common approach is to use gradient ascent.
- The \( i \)th component of the vector gradient has the form:

\[
\frac{\partial}{\partial w_i} l(w) = \sum_l x_i^l (y^l - \hat{P}(y^l = 1| x^l, w))
\]
Fitting LR by Gradient Ascent

Given this formula for the derivative of each $w_i$, we can use standard gradient ascent to optimize the weights $w$. Beginning with initial weights of zero, we repeatedly update the weights in the direction of the gradient, changing the $i$th weight according to

$$w_i \leftarrow w_i + \eta \sum_l x_i^l (y^l - \hat{P}(y^l = 1|x^l, w))$$
Regularization in Logistic Regression

- Overfitting the training data is a problem that can arise in Logistic Regression, especially when data has very high dimensions and is sparse.

\[ w = \arg \max_w \sum_l \ln P(y^l | x^l, w) - \frac{\lambda}{2} \| w \|^2 \]

- One approach to reducing overfitting is regularization, in which we create a modified “penalized log likelihood function,” which penalizes large values of \( w \).
Regularization in Logistic Regression

- The derivative of this penalized log likelihood function is similar to our earlier derivative, with one additional penalty term

\[
\frac{\partial}{\partial w_i} l(w) = \sum_l x_i^l (y_i^l - \hat{P}(y_i^l = 1| x_i^l, w)) - \lambda w_i
\]

- which gives us the modified gradient descent rule

\[
w_i \leftarrow w_i + \eta \sum_l x_i^l (y_i^l - \hat{P}(y_i^l = 1| x_i^l, w)) - \eta \lambda w_i
\]
Summary of Logistic Regression

- Learns the Conditional Probability Distribution $P(y \mid x)$
- Local Search.
  - Begins with initial weight vector.
  - Modifies it iteratively to maximize an objective function.
  - The objective function is the conditional log likelihood of the data – so the algorithm seeks the probability distribution $P(y \mid x)$ that is most likely given the data.
What you should know LR

- In general, NB and LR make different assumptions
  - NB: Features independent given class -> assumption on \( P(X | Y) \)
  - LR: Functional form of \( P(Y | X) \), no assumption on \( P(X | Y) \)
- LR is a linear classifier
  - decision rule is a hyperplane
- LR optimized by conditional likelihood
  - no closed-form solution
  - concave -> global optimum with gradient ascent