Classification & Regression

Classification (Logistic Regression)
Visualize a *Linear* Regression Line

![Linear Regression Line Graph](image)
Limits of Linear Regression

Is this meaningful?

Remember linear regression MUST go through $\bar{x}$ and $\bar{y}$
Changing How We Think about Outcomes

• Rather than predicting a target “value,” in the binary case, it makes more sense to consider *probability* of having the outcome occur at a specific X (or set of X’s)

\[ \hat{y} = \beta_0 + \beta_1 x_1 \]

\[ p(y) = \beta_0 + \beta_1 x_1 \]
Changing How We Think about Outcomes

• We know probability is bounded \([0,1]\): which makes expressions such as this difficult to solve \(p(y) = \beta_0 + \beta_1 x_1\)
  • How can we turn a proportion into a continuous outcome?
Changing How We Think about Outcomes

• We know probability is bounded $[0,1]$: which makes expressions such as this difficult to solve $p(y) = \beta_0 + \beta_1 x_1$
  • How can we turn a proportion into a continuous outcome?

• From the statistics section we remember
  – Odds of an event are defined as: $\frac{P(event)}{1-P(event)}$, which is bounded $[0, \infty]$
Changing How We Think about Outcomes

- We know probability is bounded $[0,1]$: which makes expressions such as this difficult to solve $p(y) = \beta_0 + \beta_1 x_1$
  - *How can we turn a proportion into a continuous outcome?*

- From the statistics section we remember
  - Odds of an event are defined as: $\frac{P(event)}{1-P(event)}$, which is bounded $[0, \infty]$
  - With a little math we know that:
    - Log Odds: $(\ln(\frac{P(event)}{1-P(event)})$ is bounded $[-\infty, \infty]$

  *Now we're getting somewhere*
Logistic Regression

Better idea: Set the log(odds) to the linear function.

$$\log(\text{odds}) = \logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$
Logistic Regression

Better idea: Set the log(odds) to the linear function.

\[
\log(\text{odds}) = \logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x
\]

*We are fitting a linear model in the logit scale*
Logistic Regression

\[ \log(\text{odds}) = \logit(p) = \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 x \]

- If we then, solve for \( p \) get **logistic (logit) function**; back in the correct range of values \([0, 1]\).
  - Because we really want the probability the event will occur given some \( X \)'s:

\[
p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}
\]
Visualizing the Logistic Curve

If we plot this function we find a sigmoid function that assumes values in range $[0,1]$

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$
Visualizing the Logistic Curve

![Logistic Curve Diagram]
What Distributions Can be Fit?
Logistic Regression Fits

**Simple Pattern**
- Works Perfectly

**Simple Pattern with Noise**
- Useful, some error

**Complex Pattern**
- Not useful, cannot represent larger trend
Under the Hood
Logistic Regression Optimization

• So it’s clear the form \( \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 x \) is useful but...
  – How do solve for the \( \beta \) coefficients in \( p(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}} \)?

• Unlike least squares methods, finding a closed form for the coefficients using is not possible
  – Rather, \( \beta \) coefficients estimated using technique called Maximum likelihood estimation (MLE)
MLE Intuition

• Although for this class (and most real world settings) you will use software to estimate these coefficient values, it is important to have an understanding of what MLE is doing.

• Think back to statistics notes on binomial distribution

Flipping a coin: \( P(Y = y) = \binom{n}{y}p^y(1 - p)^{n-y} \)

<table>
<thead>
<tr>
<th>Y</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(Y=y)</td>
<td>.008</td>
<td>.055</td>
<td>.0164</td>
<td>.273</td>
<td>.273</td>
<td>.164</td>
<td>.055</td>
<td>.008</td>
</tr>
</tbody>
</table>
MLE Intuition

Consider a clinical trial in which 35 independent patients are given a new medication for pain relief. 22 patients report “significant” relief 1-hr after medication.

\[ P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \]
MLE Intuition

• Consider a clinical trial in which 35 independent patients are given a new medication for pain relief.
  – 22 patients report “significant” relief 1-hr after medication
  \[ P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \]

• Break this down:
  – \( Y = \) outcome (22 patient)
  – \( N: 35 \) total patients
  – \( P\)-unknown
MLE Intuition

- Consider a clinical trial in which 35 independent patients are given a new medication for pain relief.
  - 22 patients report “significant” relief 1-hr after medication
    \[ P(Y = 22) = \binom{35}{22} p^{22} (1 - p)^{35-22} \]

- Break this down:
  - Y = outcome (22 patient)
  - N: 35 total patients
  - P-unknown
MLE Intuition

• Now for each data-point, we have a vector of features, \( x_i \), and a class, \( y_i \).
• We then want to know how likely is obtaining each of the observations
  – Independent observations so we compute the product of their probabilities
  – \( \prod_{i=1}^{n} \Pr(y_i \mid \beta + \beta_1 + \beta_2 + \cdots) \)
Math Ahead
MLE

\[ L = \prod_{i=1}^{n} \Pr(y_i) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{n-y_i} = \prod_{i=1}^{n} \left( \frac{p_i}{1-p_i} \right)^{y_i} (1 - p) \]
MLE

- \( L = \prod_{i=1}^{n} \Pr(y_i) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{n-y_i} = \prod_{i=1}^{n} \left( \frac{p_i}{1-p_i} \right)^{y_i} (1 - p) \)

- \( \ln(L) = \sum y_i \ln\left( \frac{p_i}{1-p_i} \right) + \sum \ln(1 - p) \)
  
  \text{Remember: } \ln\left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 x

- \( \text{LL} = \sum y_i \ln(\beta x_i) + \sum \ln(1 - e^{\beta x_i}) \)
  
  \text{Maximize}
MLE

• No closed form solutions:
  – Use an optimization method like Newton-Raphson

• Iterative process used, beginning with tentative solution, revises it slightly to improve, and repeats revision until “converged” (no improvement)
MLE Outcomes

• MLE chooses values for parameter estimates which make the observed data “maximally likely.”
  – These are our $\beta$’s

• **Standard errors** are obtained as a by-product of the maximization process

• What are standard errors useful for?
MLE Outcomes

• MLE chooses values for parameter estimates which make the observed data “maximally likely.”
  – These are our $\beta$’s

• Standard errors are obtained as a by-product of the maximization process

• What are standard errors useful for?
  – *Inference around the coefficients!*
WORKING ASSUMPTIONS
Logistic Regression Assumptions

• Just like linear regression, logistic regression requires the data to fulfill some assumptions to provide reliable estimates:
  – \( Y_i \) are from Bernoulli or binomial \((n, \mu)\) distribution
  – There exists an (approximately) linear relation between each \( X \) and the \text{LOGIT} of \( Y \)
  – Instances are independent of each other
  – No Multicollinearity (correlation amongst independent features \( X \))
    • As with linear regression: Failure to do so, results in instability with respect to the coefficients and their confidence intervals
“Linearity”

- Remember: There exists an (approximately) linear relation between each X and the \text{LOGIT} of Y
  - This is fairly difficult to assess, with dichotomous data
“Linearity”

• Remember: There exists an (approximately) linear relation between each $X$ and the $\text{LOGIT}$ of $Y$
  – This is fairly difficult to assess, with dichotomous data

• What we can do it fit a smoothing curve (loess)
“Linearity”

- Remember: There exists an (approximately) linear relation between each X and the LOGIT of Y
  - This is fairly difficult to assess, with dichotomous data

- What we can do it fit a smoothing curve (loess)
  - We then can take the logit of these probabilities. Straight lines are preferred here 😊
Independence / Multicollinearity

• As with linear regression, we must have a general understanding of how the data was collected
  – Our residual plots here may help with this

• Multicollinearity:
  – We can again use the bi-variate correlations and VIF metrics for all features
Logistic Model Diagnostics

• As there is no “True” value, only a binary indicator, the notion of residuals remains an open topic today

• Two Primary Residuals are often used for model diagnosis
  – Deviance
  – Pearson's
Diagnostics and Plots

Although there is no clear cut guidelines as there are with linear regression, it may still be helpful

- Deviance – Good for: identifying potential outliers
  - You can plot the deviance of each point against its index value
- Index’s have no logical meaning and the plot
Understanding Logistic Regression
Coefficients (Binary/Dummy Features)

- Thinking back to linear regression....
  - Coefficients are difference between levels and reference category of a feature

\[
\beta_1 = (\beta_0 + \beta_1 - \beta_0) = \ln 1 - \ln 0 = \ln (p_0 1 - p_0 0) - \ln (p_0 1 - p_0 0)
\]
Coefficients (Binary/Dummy Features)

- Thinking back to linear regression....
  - Coefficients are difference between levels and reference category of a feature

- Let’s compare to logit function: \( \ln \left( \frac{p(x)}{1-p(x)} \right) = \beta_0 + \beta_1 x \)
  - Plug in:
    - \( x:0 = \ln(0) = \beta_0 + \beta_1(0) = \beta_0 \)
    - \( x:1 = \ln(1) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \)
Coefficients (Binary/Dummy Features)

- Thinking back to linear regression....
  - Coefficients are difference between levels and reference category of a feature
- Let’s compare to logit function: \( \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x \)
  - Plug in:
    - \( X:0 = \ln(0) = \beta_0 + \beta_1(0) = \beta_0 \)
    - \( X:1 = \ln(1) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \)
  - Solve for \( \beta_1 = (\beta_0 + \beta_1 - \beta_0) = \ln(1) - \ln(0) \)
    - \( \beta_1 = \ln(1) - \ln(0) = \ln\left(\frac{p(1)}{1-p(1)}\right) - \ln\left(\frac{p(0)}{1-p(0)}\right) = \ln\left(\frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}}\right) \)
Coefficients (Binary/Dummy Features)

- Thinking back to linear regression....
  - Coefficients are difference between levels and reference category of a feature

- Let's compare to logit function: \( \ln \left( \frac{p(x)}{1-p(x)} \right) = \beta_0 + \beta_1 x \)
  - **Plug in:**
    - \( x:0 = \ln(0) = \beta_0 + \beta_1 (0) = \beta_0 \)
    - \( x:1 = \ln(1) = \beta_0 + \beta_1 (1) = \beta_0 + \beta_1 \)
  - Solve for \( \beta_1 = (\beta_0 + \beta_1 - \beta_0) = Ln(1) - \ln(0) \)
    - \( \beta_1 = Ln(1) - \ln(0) = \ln \left( \frac{p(1)}{1-p(1)} \right) - \ln \left( \frac{p(0)}{1-p(0)} \right) = \ln \left( \frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}} \right) \)

\( \beta_1 = \text{Odds Ratio} \)

**What is this?**
- The odds exposed: \( \frac{p(1)}{1-p(1)} \)
- The odds not exposed: \( \frac{p(0)}{1-p(0)} \)
Coefficients (Binary/Dummy Features)

• Thinking back to linear regression....
  – Coefficients are difference between levels and reference category of a feature

• Let’s compare to logit function: \( \ln \left( \frac{p(x)}{1-p(x)} \right) = \beta_0 + \beta_1 x \)
  – Plug in:
    • \( x:0 = \ln(0) = \beta_0 + \beta_1(0) = \beta_0 \)
    • \( x:1 = \ln(1) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \)

  – Solve for \( \beta_1 = (\beta_0 + \beta_1 - \beta_0) = \ln(1) - \ln(0) \)
  \( \beta_1 = \ln(1) - \ln(0) = \ln \left( \frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}} \right) \)

  – Odds Ratios! \( \beta_1 = \ln(\text{Odds Ratio}) \)
Next Class – Interpretation and Inference