Generalized Linear Model (GLM)

- Let $y_i$ have any probability distribution so long as it is from the “exponential” family
  - e.g., normal, log-normal, exponential, gamma, chi-squared, beta, Bernoulli, Poisson, etc.
  - Not included are Student’s t, mixed distributions
- Allow for any transformation of $y$ (the link function)
  - Must be monotonic and differentiable
- Linear in the model parameters

$$ g(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots $$

Non-Normal Distributions

- For a non-normal distribution, Least Squares ≠ MLE
- Combine any link function with any (exponential family) probability distribution for $y$, then find the maximum likelihood estimates for the parameters
  - Solve with iteratively reweighted least squares
  - Many software packages can do this regression

Typical Distribution/Link Pairings

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Typical uses</th>
<th>Link Name</th>
<th>Link function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Linear-response data</td>
<td>Identity</td>
<td>$g(y) = y$</td>
</tr>
<tr>
<td>Exponential or Gamma</td>
<td>Exponential-response data, scale parameters</td>
<td>Inverse</td>
<td>$g(y) = 1/y$</td>
</tr>
<tr>
<td>Poisson</td>
<td>count of occurrences in fixed amount of time/space</td>
<td>Log</td>
<td>$g(y) = \ln(y)$</td>
</tr>
<tr>
<td>Bernoulli or Binomial</td>
<td>outcome of single yes/no occurrence</td>
<td>Logit</td>
<td>$g(y) = \ln\left(\frac{y}{1-y}\right)$</td>
</tr>
</tbody>
</table>

What if our Response is Binary?

- Sometimes the response is binary
  - The patient lives or dies
  - The part passes or fails
  - The customer buys or doesn’t buy
- The response $y$ will follow a Bernoulli distribution
  - $E[y] = \pi$, the probability of “success” ($y = 1$)
  - $\text{var}[y] = \pi(1 - \pi)$ (a function of the mean)
- We want to model the probability of success
  - $\hat{y} = E[y] = \pi$

Predicting Proportions

- We want to predict a proportion (or probability), $\pi$, for a categorical variable
  - Fraction of people that die of a heart attack
  - Fraction of molecules that decompose
- Consider the following linear model

$$ \hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots $$

- Problem: $\pi$ is constrained to between 0 and 1
  - This model does not force a constraint on $\hat{y}$
  - Additionally, the variance is not constant
Probit Regression

- Instead of a linear model, assume $p$ is sigmoidally shaped
  - Note that virtually every cdf (cumulative distribution function) has a sigmoidal shape
- Example: Probit model
  - Assume $p$ is Bernoulli distributed, then

\[
\text{probit}(\pi) = \text{NormInv}(\pi) = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots
\]

Logistic Distribution

- The logistic cdf can be analytically inverted

\[
cdf = \frac{1}{1 + e^{-(x-\mu)/s}} \quad \ln \left( \frac{cdf}{1 - cdf} \right) = \frac{x - \mu}{s}
\]
- We'll model $\pi$, our output fraction of a binary variable, with this cdf
- This is called logistic regression

\[
\text{logit}(\pi) = \ln \left( \frac{\pi}{1 - \pi} \right) = \text{link function}
\]

Logistic Regression

- Instead of the probit model's inverse normal function, use the inverse logistic cdf (and assume $\pi$ is Bernoulli distributed)

\[
\text{logit}(\pi) = \ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots
\]
- Note: \( \text{odds} = \frac{\text{probability of success}}{\text{probability of failure}} = \frac{\pi}{1 - \pi} \)
- Inverting, $\pi(x_{11}, \ldots) = \frac{e^{\beta_0 + \beta_1 x_{11} + \cdots}}{1 + e^{\beta_0 + \beta_1 x_{11} + \cdots}}$

Uses of Logistic Regression

- Predict binary outcome events
  - Patient mortality after surgery (as a function of age, prior health indicators, sex)
  - Probability of contracting a certain disease (as a function of age, ethnicity, fitness, sex)
  - Probability customer will make a purchase (as a function of income, age, sex, neighborhood)
  - Probability of defaulting on a mortgage (as a function of price, income, interest rate, mortgage type)

Lecture 58: What have we learned?

- What are the three requirements of a generalized linear model?
- What are some common distribution/link function pairs?
- What is the distribution/link function pair for logistic regression?
- Name three examples of where you might want to use a logistic regression