Logistic Regression Summary Sheet

1 Theoretical Framework

Logistic Regression is used to model Binomial and Bernoulli data. Recall the derivation of the logistic link function as a result of the canonical exponential family parameterization and link function of both the bernoulli and binomial probability density functions.

$$X \sim Bern(p) \Rightarrow f_X(x|p) = \exp\left\{ \overbrace{x}^{t(x)} \overbrace{\ln(\frac{p}{1-p})}^{\theta} + \overbrace{\ln(1-p)}^{-b(\theta)} \right\}$$

$$X \sim Bin(n,p) \Rightarrow f_X(x|p) = \exp\left\{ \underbrace{x}_{t(x)} \underbrace{\ln(\frac{p}{1-p})}_{\theta} + \underbrace{n\ln(1-p)}_{-b(\theta)} + \underbrace{\ln(\binom{n}{x})}_{c(x)} \right\}$$

This leads to the following mean, variance, and link functions

	$b(\theta)$	$b'(\theta) = E[X]$	$g(\mu)$	$a(\phi)^1 b''(\theta)$
Bernoulli	$\ln(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta} = p$	$\ln(\frac{\mu_i}{1-\mu_i}) = x_i^T\beta$	$\frac{e^\theta}{(1+e^\theta)^2}$
Binomial	$n\ln(1+e^\theta)$	$\frac{ne^\theta}{1+e^\theta} = np$	$\ln(\frac{\mu_i}{1-\mu_i}) = x_i^T\beta$	$\frac{e^\theta}{(1+e^\theta)^2}$

Using the Canonical Link allows us to arrive at the following Score Function and Information Matrix

Bernoulli

$$U(\vec{\beta}) = X^T(Y-\mu) = X^T(Y-\vec{p})$$
$$I(\beta) = J(\beta) = X^TVX$$
$$V = diag(v(\mu_i)) = diag(\mu_i(1-\mu_i))$$

Binomial

$$U(\vec{\beta}) = X^T(Y-\mu) = X^T(Y-n\vec{p})$$
$$I(\beta) = J(\beta) = X^TVX$$
$$V = diag(v(\mu_i)) = diag(n_i\mu_i(1-\mu_i))$$

1.1 Score, Information Functions Derivations

The above Score and Information matrix derivations are below

$$Y_1, ..., Y_n \overset{i.i.d}{\sim} \text{Bernoulli}$$

$$\begin{aligned}
L(p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\
l(p_i) &= \ln(p_i)y_i + (1-y_i)\ln(1-p_i) \\
U(\vec{\beta}) &= \frac{\partial l}{\partial p_i}\frac{\partial p_i}{\partial \vec{\beta}} \\
\frac{\partial l}{\partial p_i} &= \frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \\
\frac{\partial p_i}{\partial \vec{\beta}} &= \vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2} \\
U(\vec{\beta}) &= \Big(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\Big)\Big(\vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2}\Big) \\
U(\vec{\beta}) &= \Big(\frac{y_i - p_i}{p_i(1-p_i)}\Big)\Big(\vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2}\Big) \\
U(\vec{\beta}) &= \Big((y_i - p_i)\frac{(1+e^{\vec{x}_i^T\beta})^2}{e^{x_i^T\beta}})\Big)\Big(\vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2}\Big) \\
U(\vec{\beta}) &= (y_i - p_i)\vec{x}
\end{aligned}$$

$$Y_1, ..., Y_n \overset{i.i.d}{\sim} \text{Binomial}$$

$$\begin{aligned}
L(p_i) &= p_i^{y_i}(1-p_i)^{n_i-y_i} \\
l(p_i) &= \ln(p_i)y_i + (n_i-y_i)\ln(1-p_i) \\
U(\vec{\beta}) &= \frac{\partial l}{\partial p_i}\frac{\partial p_i}{\partial \vec{\beta}} \\
\frac{\partial l}{\partial p_i} &= \frac{y_i}{p_i} - \frac{n-y_i}{1-p_i} \\
\frac{\partial p_i}{\partial \vec{\beta}} &= \vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2} \\
U(\vec{\beta}) &= \Big(\frac{y_i}{p_i} - \frac{n_i-y_i}{1-p_i}\Big)\Big(\vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2}\Big) \\
U(\vec{\beta}) &= \Big(\frac{y_i - n_ip_i}{p_i(1-p_i)}\Big)\Big(\vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2}\Big) \\
U(\vec{\beta}) &= \Big((y_i - n_ip_i)\frac{(1+e^{\vec{x}_i^T\beta})^2}{e^{x_i^T\beta}})\Big)\Big(\vec{x}_i \frac{e^{\vec{x}_i^T\beta}}{(1+e^{\vec{x}_i^T\beta})^2}\Big) \\
U(\vec{\beta}) &= (y_i - n_ip_i)\vec{x}
\end{aligned}$$

Note that for the vector of $Y = (y_1, ..., y_n)$ this extends to

$$U(\vec{\beta}) = X^T(Y-\mu) = X^T(Y-\vec{p}) \qquad U(\vec{\beta}) = X^T(Y-\mu) = X^T(Y-n\vec{p})$$

This is the same form as shown above.

2 Interpretation

Log-Odds Model
This model is referred to as the log-odds model in clinical parlance, for the fact that the interpretation of the coefficients in the model is in terms of the log odds of the outcome variable occurring.

Assume the following model, where $Y_i \sim \text{Bernoulli}(\pi_i)$, $x_i \in \mathbb{R} \quad \forall i$

$$\ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i$$

Interpretations can be derived as follows

Interpretation: β_0 is the log odds of Y occuring, when $x = 0$
Implications: e^{β_0} is the odds of Y occurring when x = 0

$$\frac{e^{\beta_0}}{1 + e^{\beta_0}} = P(Y = 1|X = 0)$$

Justification:

$$\ln(\frac{\pi_i}{1 - \pi_i}) = \beta_0 \iff x_i = 0$$

Interpretation: β_1 is the log odds ratio of Y occuring for each unit change in X. The log odds ratio may also be understood as the difference in log odds.

Justification:

$$\text{log odds }(Y = 1|X = x + 1) - \text{log odds }(Y = 1|x = x + 1) = \beta_0 + \beta_1(x + 1) - \beta_0 - \beta_1 x = \beta_1$$

$$\ln(\frac{\pi_i}{1 - \pi_i}) - \ln(\frac{\pi_j}{1 - \pi_j}) = \beta_1$$

$$\ln\left(\frac{\pi_i/(1 - \pi_j)}{(1 - \pi_i)\pi_j}\right) = \beta_1$$

3 Parameter Estimation

3.1 Saturated Model

A **saturated model** is defined as having as many parameters as there are observations. For a 2 × 2 contigency table, this implies that estimates of β_0, β_1 may be derived analytically.

Reference Table

		Health Status		
		$Y = 0$	$Y = 1$	total
Treatment	$X = 0$	n_{00}	n_{01}	n_{0+}
	$X = 1$	n_{10}	n_{11}	n_{1+}
	total	n_{+0}	n_{+1}	n

The model is given by :

$$\ln(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_1 x_i$$

Parameter Estimates:

- β_0: log odds when x = 0 $\Rightarrow \hat{\beta}_0 = \frac{n_{01}/n_{0+}}{n_{00}/n_{0+}} = \ln(\frac{n_{01}}{n_{00}})$
- $\beta_0 + \beta_1$: log odds when x =1

  $$\hat{\beta}_1 + \hat{\beta}_0 = \ln(\tfrac{n_{11}/n_{1+}}{n_{10}/n_{1+}}) = \ln(\tfrac{n_{11}}{n_{10}})$$

- $\hat{\beta}_1 = \ln(\frac{n_{11}}{n_{10}}) - \ln(\frac{n_{01}}{n_{00}}) = \ln(\frac{n_{11}n_{00}}{n_{10}n_{01}})$

Note that these parameters could also be derived by *maximum likelihood estimation* using the following model parameterization

$$\begin{aligned}
\ln(\frac{\pi_i}{1-\pi_i}) &= \alpha(1-x_i) + \beta x_i \\
\Rightarrow P(Y_i = 1|X_i = 0) = p_1 &= \frac{e^\alpha}{1+e^\alpha} \iff P(Y_i = 0|X_i = 0) = 1 - p_1 = \frac{1}{1+e^\alpha} \\
P(Y_i = 1|X_i = 0) = p_2 &= \frac{e^\beta}{1+e^\beta} \iff P(Y_i = 0|X_i = 0) = 1 - p_2 \frac{1}{1+e^\beta} \\
Y_1, ..., Y_n &\overset{i.i.d.}{\sim} \text{Bernoulli}(p_k) \quad k = 1, 2 \\
L(p) &= p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} \\
L(p) &= p^{\sum_{i=1}^n y_i|x_i=1} p^{\sum_{i=1}^n y_i|x_i=0}(1-p)^{n-\sum_{i=1}^n y_i|x_i=1}(1-p)^{n-\sum_{i=1}^n y_i|x_i=0} \\
L(\alpha, \beta) &= \left(\frac{e^\alpha}{1+e^\alpha}\right)^{n_{01}} \left(\frac{e^\beta}{1+e^\beta}\right)^{n_{11}} \left(\frac{1}{1+e^\alpha}\right)^{n_{00}} \left(\frac{1}{1+e^\beta}\right)^{n10} \\
l(\alpha, \beta) &= n_{01}\alpha - (n_{01} + n_{00})\ln(1+e^\alpha) + n_{11}\beta - (n_{10} + n_{11})\ln(1+e^\beta) \\
\Rightarrow \frac{\partial l}{\partial \alpha} &= n_{10} - \frac{(n_{01} + n_{00})e^\alpha}{1+e^\alpha} = 0 \iff \alpha = \ln(\frac{n'}{1-n'}) \quad n' = \frac{n_{10}}{n_{01} + n_{00}} \\
\Rightarrow \frac{\partial l}{\partial \beta} &= n_{11} - \frac{(n_{10} + n_{11})e^\beta}{1+e^\beta} = 0 \iff \beta = \ln(\frac{n^*}{1-n^*}) \quad n^* = \frac{n_{11}}{n_{10} + n_{11}}
\end{aligned}$$

4 Hypothesis Testing

4.1 Wald Test

- General Form: $H_0 : C\beta = \vec{d}$ vs. $H_a : C\beta \neq d$

$$\text{Test statistic } = (C\hat{\beta} - d)^T\{CI^{-1}(\hat{\beta})C^T\}^{-1}(C\hat{\beta} - d) \sim \chi^2_r$$

  note that $\text{rank}(C) = r$

- if testing $H_0 : \beta_j = 0, H_a : \beta_j \neq 0$

$$\text{test statistic } = \frac{\hat{\beta}_j^2}{\hat{SE}(\hat{\beta}_j)^2} \sim \chi^2_1$$

4.2 Likelihood Ratio Test

- Available through difference of Deviances or differnece of the log likelihood ratio using different $\hat{\beta}$ estimates

$$TS = \{2l(\hat{\beta}^0) - l(\hat{\beta})\} \sim \chi^2_{\dim(\hat{\beta})-\dim(\hat{\beta}^0)}$$
$$TS = D_0 - D_1 \sim \chi^2_{\dim(\hat{\beta})-\dim(\hat{\beta}^0)}$$

5 Goodness of Fit Testing

5.1 Residuals

- Pearson residuals
  $\hat{r}_i^P = \frac{Y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}$

- Deviance residuals

$$\hat{r}_i^D = \text{sign}(Y_i - n_i\hat{p}_i)\sqrt{|D_i|}$$

- Deviance Test

$$D = 2\sum_{j=1}^{n}\left[Y_i \log(\frac{Y_i}{n_i\hat{p}_i}) + (n_i - Y_i)\log(\frac{n_i - Y_i}{n_i - n_i\hat{p}_i})\right] \sim \chi^2_{n-q}$$

- Pearson χ^2 Test

$$X_p^2 = \sum_{i=1}^{n} \frac{(Y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} \sim \chi^2_{n-q}$$

- Deviance, Pearson χ^2 work well when expected number of events and non-events > 5
- the more events the better

5.2 Hosmer-Lemeshow Test

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g p_g(1-p_g)} \sim \chi^2_{G-2}$$

Algorithm 1 Hosmer Lemeshow Test

Sort fitted p_i's
split p_i's into some G number of groups (usually $G = 10$)
for $g = 1$ to G **do**
  $O_g = p$[all i in group g] 'The observed number of events'
  p_g =mean(p[all i in group g]) 'The average of the p_i's in group g
  N_g = count(number of p_i's in group g)
  $E_g = N_g * p_g$
end for
return $\sum_{g=1}^{G} \frac{(O_g-E_g)^2}{N_g p_g(1-p_g)} \sim \chi^2_{G-2}$

5.3 R^2

$$R^2 = 1 - \left\{ \frac{L(\hat{p}_i{}^{interceptmodel})}{L(\hat{p})} \right\}^{\frac{2}{N}}$$

Max Adjusted R^2

$$R^2_{MA} = \frac{R^2}{\max(R^2)}$$