Logistic regression : Interpreting model outputs

Logistic regression The basic equations of a logistic regression model where there are k predictor variables $x_1, \ldots, x_k$ and a binary target variable (levels coded as 1 and 0) are as below.

- $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

$$\sigma(z) = \frac{1}{1 + e^{-z}} = p$$

where $\sigma(z)$ is the sigmoid function.

$$z = \sigma^{-1}(p) = ln\left(\frac{p}{1-p}\right)$$

$ln\left(\frac{p}{1-p}\right)$ is referred to as the logit function.

1 Interpreting model outputs of logistic regression

For a short and readable description of output of glm() function in R, you may refer the link `https://www.theanalysisfactor.com/r-glm-model-fit/`
A more detailed, yet readable description with examples can be found at `http://www.stat.columbia.edu/~martin/W2024/R11.pdf`

1.1 Null and residual deviances

Deviance is a measure of "badness of fit". Higher the deviance values, worse is the fit. A brief summary is outlined below.

- The null deviance shows how well the response is predicted by the model with nothing but an intercept.
- The residual deviance shows how well the response is predicted by the model when the predictors are included.

The residual deviance can be used to test for goodness of fit of the model, in R this can be done by
p-value = 1 - pchisq(deviance, degrees of freedom)

A slightly more detailed description is given below

- The Saturated Model is a model that assumes each data point has its own parameters (which means you have n parameters to estimate.)
- The Null Model assumes the exact "opposite", it assumes one parameter for all of the data points, which means you only estimate 1 parameter.
- The Proposed Model assumes you can explain your data points with p parameters + an intercept term, so you have p+1 parameters.
- If the Null Deviance is really small, it means that the Null Model explains the data pretty well, likewise with the Residual Deviance.
- $NullDeviance = 2(LL(SaturatedModel) - LL(NullModel))$ on df $= df_{Sat} - df_{Null}$
- Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model)) on df = $df_{Sat} - df_{Proposed}$

  where $LL()$ represents log-likelihood.
- If the proposed model is "good" then the Deviance approximately follows a χ^2 distribution with $(df_{sat} - df_{model})$ degrees of freedom.
- To compare the Null model with the proposed model, one can look at (Null Deviance - Residual Deviance) which would approximately follow a χ^2 distribution with degrees of freedom = dfProposed - dfNull = (n-(p+1))-(n-1)=p

Interpreting model outputs of logistic regression
Assume there are n data points and a model has p parameters.
Note : Degrees of freedom unless specified otherwise $= n - p$ **Null and residual deviances**
Deviance is a measure of "badness of fit". Higher the deviance values, worse is the fit.

- The null deviance shows how well the response is predicted by the model with nothing but an intercept.
- The residual deviance shows how well the response is predicted by the model when the predictors are included.

The residual deviance can be used to test for goodness of fit of the model, in R this can be done by
p-value = 1 - pchisq(deviance, degrees of freedom)

Summary of various outputs

- The Saturated Model is a model that assumes each data point has its own parameters (which means you have n parameters to estimate.)
- The Null Model assumes the exact "opposite", in that is assumes one parameter for all of the data points, which means you only estimate 1 parameter.
- The Proposed Model assumes you can explain your data points with p parameters + an intercept term, so you have p+1 parameters.
- If your Null Deviance is really small, it means that the Null Model explains the data pretty well. Likewise with your Residual Deviance.
- $NullDeviance = 2(LL(SaturatedModel) - LL(NullModel))on df = df_{Sat} - df_{Null} Residual Deviance = 2(LL(SaturatedModel) - LL(ProposedModel))df = df_{Sat} - df_{Proposed}$
  where $LL()$ represents log-likelihood.
- If the proposed model is "good" then the Deviance is approximately $\sim \chi^2$ with $(df_{sat} - df_{model})$ degrees of freedom.
- To compare you Null model with your Proposed model, one you can look at
  (Null Deviance - Residual Deviance) $\sim \chi^2$ with df Proposed - df Null = (n-(p+1))-(n-1)=p