Binary Logistic Regression

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Generalized linear models (GLMs) can be used in more situations than linear regression
    - More flexibility in the response variable.
    - Relaxation of some of the assumptions required.
- One such case is when there is a binary response variable.

Binary Response Variables

- **Binary response variables** often take the form of 1 (yes) or 0 (no).
- Binary response variables may be found in many fields.
- The *binomial distribution* is a member of the exponential family and is often used to model binary outcomes.

Review: Binomial Distribution I

- The **binomial distribution** models n independent Bernoulli trials each with the probability p of *success*.
- Example: Probability of a coin flipped $n = 10$ times landing on tails 7 times.

Review: Binomial Distribution II

- The PMF of the binomial distribution is:

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{for } x = 0, 1, ..., n \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

- Expected value: np
- Variance: $np(1-p)$

Binary Logistic Regression

- **Binary logistic regression** (logistic regression) is a generalized linear model that is useful for when there is a dichotomous response variable.

- Assume that π is the probability of a success ($Y = 1$).

- Recall from GLMs:
  $g\left[E(\boldsymbol{Y}|\boldsymbol{X})\right] = g[\boldsymbol{\mu}] = \boldsymbol{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$

- **Odds ratio**: $\frac{\pi}{1-\pi}$
  - How much more likely the occurrence of the event is, compared to the non-occurrence.

Logistic Regression Model

- When constructing logistic regression, the log-odds, called the *logit*, is modeled in terms of explanatory variables:

$$g(\pi) = ln(\frac{\pi}{1-\pi}) = \boldsymbol{X}\beta \tag{2}$$

$$\pi = \frac{e^{\boldsymbol{X}\beta}}{1+e^{\boldsymbol{X}\beta}}$$

- The *logit* link ensures that all estimates of π are on the interval $(0, 1)$.

Logistic Regression in R

- Estimating a logistic regression model in R is very similar to a linear regression model:
  - `LogitModel <- glm(binary.response ~ Var.1 + Var.2 + ... + Var.j, family = binomial(link = "logit"), data = data)`
- Estimated using MLE.
- `link = "logit"` is included by default.

Example 1

- Import the *TitanicSurvival* dataset from the *carData* package into R.
- Take a moment to get to know the data.
- Estimate the following binary logistic regression model:

$$ln(\frac{\pi}{1-\pi}) = 1 + sex + age + passengerClass$$

$$where\ \pi = the\ probability\ of\ survival\ (Y = 1)$$

- What do the coefficients actually tell us?

Interpreting Coefficients

- Using the `summary()` function, the coefficients are presented in the **log odds**.
- For a one unit change in the explanatory variable, there is a corresponding change in the natural logarithm of the odds of a *success*.
- Positive values indicate an increase in probability and negative values indicate a decrease in probability.

Alternative Coefficients

- The estimated odds ratios may also be examined.
- The **odds ratio** quantifies the strength of the association between two events.
    - Odds ratio $= 1$: The events are independent (no relationship).
    - Odds ratio > 1: An increase in the explanatory variable increases the odds (probability) of a success.
    - Odds ratio < 1: An increase in the explanatory variable decreases the odds (probability) of a success.

Alternative Coefficients in R

- Recall, the default coefficients are presented in the **log odds**.
    - Can use the natural exponential function to obtain the odds ratios:
      `exp(coef(LogitModel))`
    - Including the 95% confidence interval:
      ```
      exp(cbind(OddsRatio = coef(LogitModel),
      confint(LogitModel)))
      ```

Example 2

- Using the *TitanicSurvival* data from Example 1, do the following:
  1. Add a random (normally distributed) explanatory variable X1 to the dataset.
  2. Estimate the same model from Example 1 including X1 and change the reference category of passengerClass to 2nd.
  3. Interpret the results of your estimated model.
  4. Examine the odds ratio estimates and interpret their meaning.

Assumptions

- We have relaxed the normality of the response and homoscedasticity assumptions.
- **The following assumptions still apply for binary logistic regression:**
  1. Binary response variable
  2. There is a linear relationship between the continuous predictor variables and the *logit* of the dependent variable.
  3. There is **no** multicollinearity of the explanatory variables.

Linearity IA

- We can use visualizations to verify this assumption.

- Directly plot the relationships of numeric variables (logit of the response vs explanatory variables) **after a model has been estimated**:
  - `probabilities <- predict(LogitModel, type = "response")`

  - `mydata <- data %>%`
  - `na.omit() %>%`
  - `dplyr::select_if(is.numeric)`
  - `predictors <- colnames(mydata)`

  - `mydata <- mydata %>%`
  - `mutate(logit = log(probabilities/(1-probabilities))) %>%`
  - `gather(key = "predictors", value = "predictor.value", -logit)`

Linearity IB

- Create the scatterplots:
  - `ggplot(mydata, aes(logit, predictor.value))+`
  - `geom_point(size = 0.5, alpha = 0.5) +`
  - `geom_smooth(method = "loess") +`
  - `theme_bw() +`
  - `facet_wrap(~predictors, scales = "free_y")`
- If the individual plots show an approximately linear relationship, we can say the model passes the linearity assumption.

Linearity II

- Directly plot the relationships of **individual** numeric variables with the residuals **after a model has been estimated**:
  - `data %>%`
  - `mutate(comp_res = coef(LogitModel)["variable.name"]*variable.name + residuals(LogitModel, type = "working")) %>%`
  - `ggplot(aes(x = variable.name, y = comp_res)) +`
  - `geom_point() +`
  - `geom_smooth(color = "red", method = "lm", linetype = 2, se = F) +`
  - `geom_smooth(se = F)`

- The red line is the linear fit and the blue line is the conditional mean.
- If the relationship is linear, the lines will be close to each other.

Example 3

- Use the above methods to check the linearity assumption of the logistic regression model that was estimated in Example 2.

Linearity Solutions

- If the linearity assumption is violated:
  - Try to transform explanatory variables to create a linear relationship (polynomials).
  - May be able to use regression splines.

Multicollinearity

- The no multicolinearity assumption can be checked using the `vif()` function from the *car* package.

- Recall: *If the value is larger than 5 or 10 we should consider removing one or more of the variables.*
- Examine the `GVIF^(1/(2*Df))` when there are 2 or more degrees of freedom.
  - *Square this value.*

Example 4

- Check the model estimated in Example 2 for the presence of multicolinearity.

Outliers

- **Regression outliers** are those observations whose values (of the response and explanatory variables) deviate from the regression relationship which holds for the majority of observations.

- Cook's distance may be used to examine logistic regression models for potential outliers (values over 0.5 and 1.0).

- In R:
  - Plots:
    `plot(LogitModel,3)` AND `plot(LogitModel,4)`
  - To get the numeric values:
    `cooks.distance(LogitModel)`

Example 5

- Check the model estimated in Example 2 for the presence of outliers.

Deviance I

- We would like to measure the the *fit* of the model.
- The Adjusted-R^2 is no longer applicable for GLMs.
  - We can use the deviance to measure the fit of the model.
- The deviance is based on the highest possible likelihood for the given data, link function, and assumed distribution.

Deviance II

- The highest possible likelihood is calculated using a *saturated model.*
  - A model where the number of parameters is the same as the number of observations (do not use in practice).
- **Deviance** is defined as twice the difference between the log-likelihood of the saturated model and the estimated model.

$$Deviance = 2(l_{saturated} - l_{estimated})$$

- The larger value of the deviance the *worse* the model is.
- Directly in R:
  - `deviance()`

Likelihood Ratio Test

- If we wish to compare two models we can use a likelihood ratio test.
- Recall: **We want to choose the simpler model unless the more complex model performs significantly better.**
- Generally used to test for the need to include blocks of variables.
- Likelihood ratio test in R:

  ```
  anova(glm.simple, glm.complex, test="LR")
  ```

Information Criteria

- We can use AIC and BIC to compare models.
- Remember, lower values of AIC and BIC are considered to be better.
- BIC has a higher penalization for the number of parameters than AIC.
- In R:
  - `AIC()`
  - `BIC()`

Example 6

- Estimate the following binary logistic regression models:
  1. $ln(\frac{\pi}{1-\pi}) = 1 + passengerClass$
  2. $ln(\frac{\pi}{1-\pi}) = 1 + sex + age + passengerClass$
  3. $ln(\frac{\pi}{1-\pi}) = 1 + sex + age^2 + passengerClass + X1$

  $where\ \pi = the\ probability\ of\ survival\ (Y = 1)$

- Compare the models using the techniques we have covered and select the best of these three models.

Comments on Model Selection

- You can also use stepwise selection on GLMs including logistic regression models.
- Parameters' significance is tested using a **Wald's test** instead of a t-test.
    - The interpretation is the same.
    - Also used for the confidence intervals of coefficient estimates.
- These methods are used to make inferences.

Predictions

- Predictions can be made using logistic regression models.
- To predict the probabilities (based on new data) in R:
  - `prediction <- predict(LogitModel, New.Data, type="response")`
- If you would like to assign a 1 or a 0 to your predictions:
  - `ifelse(prediction > 0.5, 1, 0)`

Confusion Matrix

- To assess the accuracy of the predictions made from a binary logistic regression model a confusion matrix can be used.
- A **confusion matrix** is a matrix used to assess the accuracy of a classification model.
  - A tabular summary of the number of correct and incorrect predictions made by a classifier.
- The correctly classified counts will be on the diagonal and the misclassified will be on the off diagonal.

Confusion Matrix Example

- This is the information provided by a confusion matrix for a binary classifier:

		Reference (Actual)	
		No	**Yes**
Prediction	**No**	True No	False Negative (Type II error)
	Yes	False Positive (Type I error)	True Yes

- In R:
  - `confusionMatrix(prediction, actual.outcome)`

Kappa

- *The **Kappa values** measure the accuracy of predictive models while accounting for an expected accuracy driven by random chance.*

- The Kappa value has a maximum value of 1 and larger values indicate better performance.

$$\kappa = \frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)}$$

- 0.21 - 0.40 fair, 0.41 - 0.60 moderate, 0.61 - 0.80 substantial, 0.81 - 1.00 almost perfect.

K-Fold Cross-Validation in R

- There are many existing functions that can be used to perform cross-validation in R.

- We can use the caret package:
  - `library(caret)`
  - `set.seed(2020)`
  - `train.control <- trainControl(method = "cv", number = K)`
  - `LogitModel <- train(Response ~ Var.1 + ...  + Var.M, data = data, method = "glm", family = "binomial", trControl = train.control)`
  - `print(LogitModel)` *Prints the results including accuracy and Kappa*

- Note: We can use the accuracy and Kappa values to compare models.

`confusionMatrix()` Results

- Sensitivity:
  - True Positive Rate (TPR) $= \frac{\text{TP}}{\text{P}}$
- Specificity:
  - True Negative Rate (TNR) $= \frac{\text{TN}}{\text{N}}$
- Pred Value:
  - The *positive predictive value* is defined as the percent of predicted positives that are actually positive while the *negative predictive value* is defined as the percent of negative positives that are actually negative.
- Prevalence:
  - How often does the *no* condition occur.
- Detection Rate:
  - How often is the *no* condition accurately predicted overall.
- Balanced Accuracy:
  - Balanced Accuracy $= \frac{\text{TPR} + \text{TNR}}{2}$

Example 7

- Compare the three models you estimated in Example 6 based on their prediction accuracy (10-fold cross-validation).
- Alter the provided code to examine the confusion matrix for each model.

Probit Regression

- The *logit* link function is the most popular link function when examining a binary response variable.
- Another link function called the *probit link function* may be used.
- It uses the cumulative normal distribution as the link function:

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \rightarrow \pi = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...)$$

- ***It is assumed that the error term follows the normal distribution.***
    - **The other assumptions of logistic regression apply.**

Probit Regression in R

- To estimate a probit model in R, just change the link function:
  - `ProbitModel <- glm(binary.response ~ Var.1 + Var.2 + ... + Var.j, family = binomial(link = "probit"), data = data)`

- *Note: you can use probit regression to make predictions like you would with logistic regression.*

Interpreting the Results of Probit Models I

- The coefficient estimates show how a one unit change in X is associated with a change in the z-score of Y.
  - Not necessarily intuitive to interpret.
- We can look at the results in terms of probability through the marginal effects.
  - How much does a one unit change in X impact the probability of a *success*.
- To examine the average marginal effects in R (sjPlot):
  - `plot_model(Model.name, type = "pred", terms = "variable.name")`

Interpreting the Results of Probit Models II

- The previous method only examines the average marginal effects.
  - In practice, the change may not be constant over the values of X.

- We can use predicted probabilities at different levels of X to examine this relationship:
  - `ggpredict(ProbitModel, terms = "variable.name[lower:upper by = step.length]") %>%`
  - `plot()`
- *This method can be used to examine the marginal effects of a logistic regression model.*

Example

- Use probit regression to estimate the following model:

  $$\Phi^{-1}(\pi) = 1 + \mathit{sex} + \mathit{age} + \mathit{passengerClass}$$

- Examine the marginal effects of the explanatory variables.
  - Numeric and visual results.
- Does the error term pass the normality assumption?

Exercise 1

- Take some time to estimate some regression models with binary response variables, run appropriate diagnostics, and make predictions.
- Note: *Non-binary response variables can be converted into binary response variables.*

  Example: Income higher or lower than national average (1=yes, 0=no)

References & Resources

1. De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

2. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276-282.

- `glm()`
- `family()`
- Categorical Regression in Stata and R
- anova()
- `AIC()`
- `confusionMatrix()`
- `trainControl()`
- `train()`
- `margins()`
- ggeffects