NYC DATA SCIENCE ACADEMY

Generalized Linear Models

Data Science Bootcamp

Outline

- Part 1: Generalized Linear Models
- Part 2: Logistic Regression
- Part 3: Maximum Likelihood Estimation
- Part 4: Model Interpretation
- Part 5: Assessing Model Fit
- Part 6: Review

PART 1

Generalized Linear Models

What are Generalized Linear Models?

- Generalized linear models are a family of models that extend the idea of ordinary least squares regression beyond the assumptions of what we have seen thus far in simple and multiple linear regression.

- In particular, a few assumptions may be relaxed:
  - Response variables need not have error distributions that are normal.
  - Response variables need not have constant variances.

- While still linear in form, the response variable is related to the linear function by the agency of a link function; the choice of link function is what generates the family of generalized linear models.

Generalized Linear Models: Mathematically

- The class of generalized linear models can be described as follows:
  1. $Y_1, Y_2, \ldots, Y_n$ are independent responses that follow a probability distribution belonging to the exponential family of probability distributions and have expected value $E[Y_i] = \mu_i$.
  2. A linear predictor based on the predictor variables $X_{i1}, X_{i2}, \ldots, X_{ip}$ is created as follows:

$$X_i'\beta = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$$

  3. The link function g relates the linear predictor to the mean response:

$$X_i'\beta = g(\mu_i)$$

- Generalized linear models may have non-constant variances σ_i^2 for the responses Y_i, but the variance must be a function of the predictor variables through the mean response μ_i.

PART 2

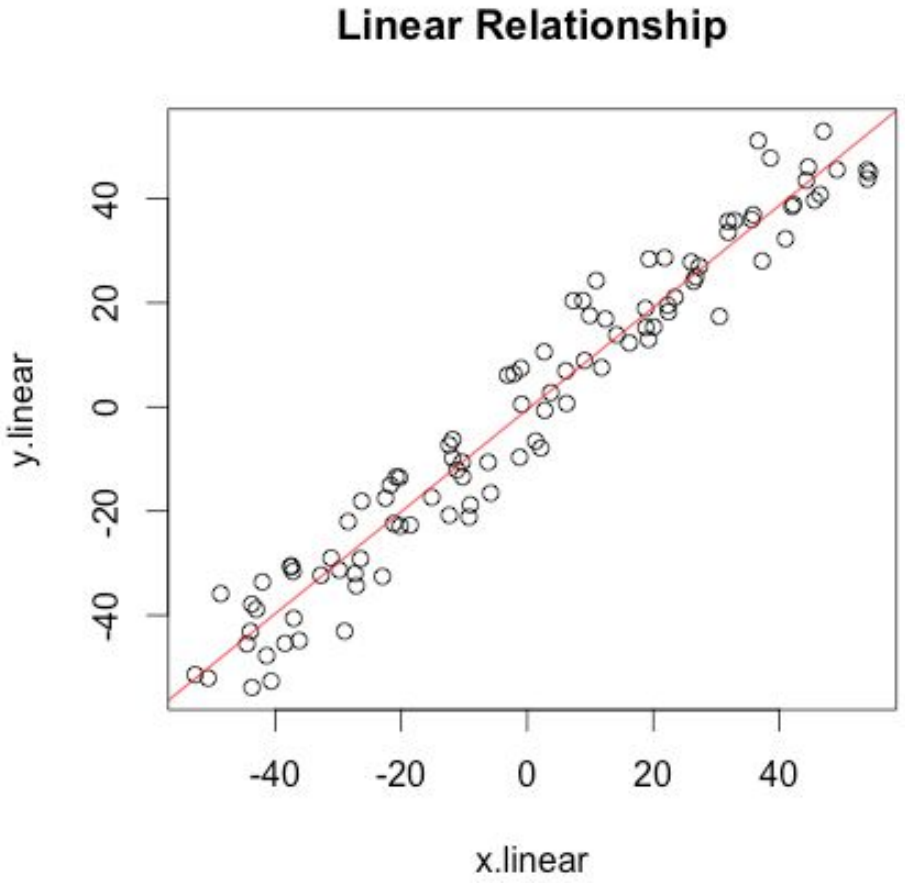
Logistic Regression

Recall: Simple Linear Regression

- Simple linear regression is a supervised machine learning method that aims to uncover a linear relationship between two continuous variables:
  - The explanatory/independent/input variable X.
  - The response/dependent/output variable Y.

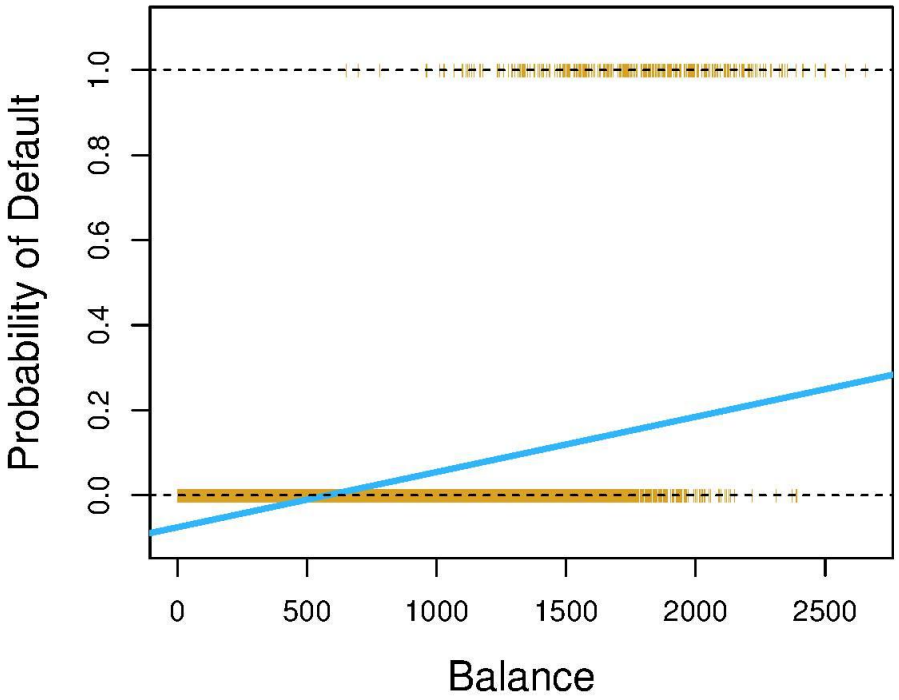
- The ultimate goal is to use this relationship to make predictions about observations not within our dataset. We answer the question:
  - If I have the value of X, what should my best guess for Y be?

Recall: Simple Linear Regression



Binary Response Variables: The Problem

- What if you have a dependent variable that is binary (i.e., there are only two choices for the outcome)?



Binary Response Variables: The Problem

- Some problems with applying simple linear regression to scenarios in which we have a binary response variable are as follows:
  - The error terms should be normally distributed around the prediction line; there should be no distinct pattern around the regression line.
  - The linear model should stray from predicting values that are impossible given our data.
  - The linear model should yield accurate predictions of the dependent variable, better than just guessing the mean -- that's the whole point of linear regression!

Binary Response Variables: The Problem

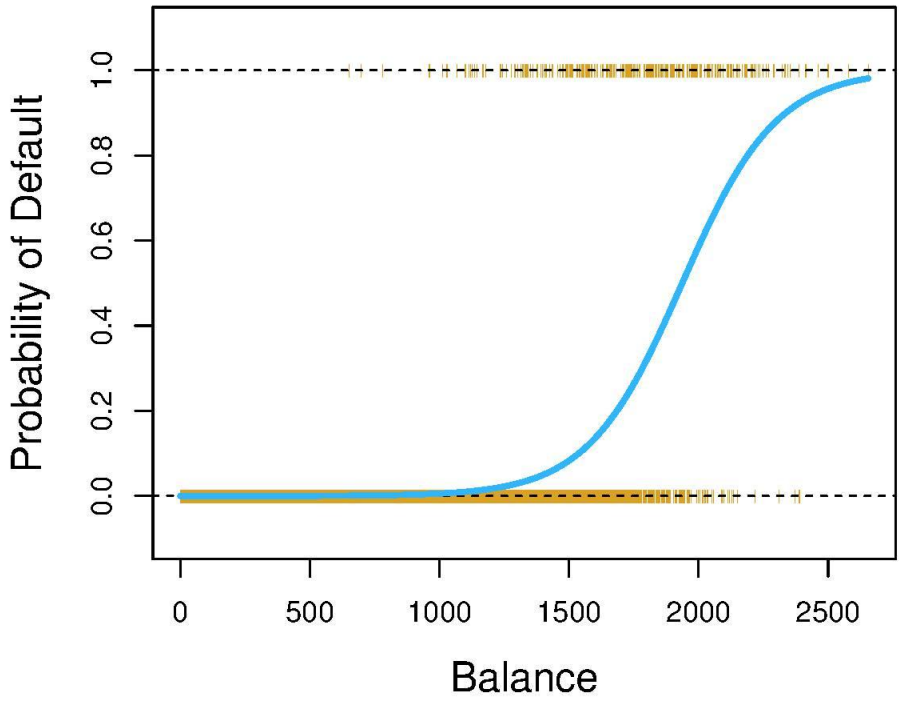
- The bottom line (pun intended): if we use linear regression to predict a binary response, our predictions would be useless; the line simply cannot fall where the majority of the data exists.

- We need a better method to create a model that tends to:
  - Hover over where the data are located.
  - Not hover over where the data are not located.

- Intuitively for this kind of data, what would the curve look like?

Binary Response Variables: The Solution

- What if we could create an S-shaped curve, i.e., a sigmoid? This would alleviate the aforementioned issues.



Binary Response Variables: The Solution

- The process of fitting a sigmoid function to binary categorical data is typically referred to as logistic regression; but how do we get an equation that follows this S-shape? We do not model the response variable directly.

- Rather than directly modeling the categories of a response variable, logistic regression models the probability that an observation belongs to a particular category.

- Maybe we can find a creative way to use a link function to connect the probability of an event to the predictor variables...

Binary Response Variables: The Solution

- Consider the case of predicting a binomial random variable Y. Recall that a binomial process is one that possesses the following properties:
  1. There are n identical trials.
  2. Each trial results in either a success S or a failure F.
  3. The probability of success p is the same for all trials; thus, the probability of failure (1 - p) is also the same for all trials.
  4. Each trial is independent of every other trial.

- In logistic regression, we wish to model p (and therefore Y) based on the predictor variables X_1, X_2, …, X_p.

Probability & Odds

- When discussing the probability p of success, we're essentially enumerating all the occurrences of success and dividing by the total number of observed events (whether they were successes or not).
  - Note that probabilities are bound by [0, 1].

- When discussing the odds of success, we're essentially considering the average number of successes we would expect to see for each failure; odds are calculated as follows:

$$Odds = \frac{p}{1-p}$$

  - Note that odds are bound by [0, ∞).

Probability & Odds

- Suppose I have a fair 6-sided die. Consider the following events:
  - Success: Rolling a 2 or 5.
  - Failure: Rolling a 1, 3, 4, or 6.

- What is the probability p of success?

$$p = \frac{\#\ Successes}{\#\ Events} = \frac{2}{6} = \frac{1}{3}$$

- What are the odds of success?

$$Odds = \frac{p}{1-p} = \frac{\frac{1}{3}}{1-\frac{1}{3}} = \frac{1}{2}$$

Log Odds as the Link Function

- What happens if we choose the link function to be the natural log of the odds ratio (also known as the log odds or logit):

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = X\beta$$

$$\frac{p}{1-p} = e^{X\beta}$$

$$p = e^{X\beta} - pe^{X\beta}$$

$$p + pe^{X\beta} = e^{X\beta}$$

$$p(1 + e^{X\beta}) = e^{X\beta}$$

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}} = \frac{1}{1 + e^{-X\beta}}$$

Log Odds as the Link Function

- Let's analyze the resulting equation a bit further:

$$p = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}$$

  - When $e^{X\beta}$ tends towards 0, p tends towards 0.
  - When $e^{X\beta}$ tends towards ∞, p tends towards 1.

- Notice: p follows the rules of basic probability by remaining bound by [0, 1]!

Log Odds as the Link Function

- Using the log odds as the link function, we are able to connect our linear model to our response in such a way that creates a sigmoidal curve to better fit our data.

- Notice that while logistic regression is a kind of nonlinear regression (because the equation for the probability p is not linear in respect to the coefficients), the nonlinearity is contained solely in the link function.
  - The regression structure that we are accustomed to from simple/multiple linear regression can still be retrieved. This is where we get the idea of generalized linear models.

PART 3

Maximum Likelihood Estimation

Estimating the Coefficients: Maximum Likelihood Estimation

- Recall that when we estimated the coefficients for linear regression we treated the scenario as a minimization problem where we chose parameter estimates to reduce the sum of squared errors.

- Although we could conduct a similar procedure and perform a nonlinear version of least squares estimation, this can get a bit complicated.
  - Instead, we will explore a new method that is often used in statistics: maximum likelihood estimation.

Estimating the Coefficients: Maximum Likelihood Estimation

- The idea behind maximum likelihood estimation:
  - Our dataset consists of actual observed outcomes; for the data we intend to use for logistic regression, the outcomes are one of two categories:
    - Success, which we will label "1".
    - Failure, which we will label as "0".
  - Using the logistic function we just derived, we can find the probability of success by passing our observations through the function:

$$p(X) = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}$$

    - Similarly, we can find the probability of failure by calculating 1 - $p(X)$.

Estimating the Coefficients: Maximum Likelihood Estimation

- The idea behind maximum likelihood estimation (continued):
  - Since the observations are assumed to be independent of one another, we can calculate the probability of our observed data by multiplying these probabilities together:

$$l(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1-p(x_j))$$

  - This equation represents the joint probability of successes and failures within our dataset; notice that the resulting equation involves the unknown parameters.

Estimating the Coefficients: Maximum Likelihood Estimation

- One set of parameter values is more likely than another set if it gives the observed outcome a higher probability of occurrence.
  - Using maximum likelihood estimation in tandem with observed outcomes allows us to hone in on the best estimates of the coefficients.

- We wish to select values of $\beta_0, \beta_1, \ldots, \beta_p$ such that the likelihood of observing our particular dataset is as high as possible.
  - If we could iterate across many combinations of parameter estimate values, we could select the combination that produces the highest likelihood for our data.

- The parameter estimates that yield the highest likelihood for our data are called the maximum likelihood estimates MLEs.

Notes on Maximum Likelihood Estimation Calculation

- We alluded to potentially using a computer routine to search through many possible parameter values to find the MLEs; however, this is unnecessary.

- As we saw with least squares estimates in simple/multiple linear regression, calculus and linear algebra provide a method by which we can arrive at these estimates in a more efficient manner.
  - Unfortunately, whereas in simple/multiple linear regression we calculated closed-form expressions for the estimates, this does not occur in logistic regression setting.
  - For logistic regression, some iterative procedures are implemented by statistical software in tandem with MLE (Newton-Raphson computation).

Properties of Maximum Likelihood Estimation

- With a sufficiently large sample size and a valid model, some properties of maximum likelihood estimation are as follows:
  1. The MLEs are essentially as unbiased as possible.
     a. Thus, the MLE principle leads to accurate estimates of the coefficients.
  2. We can derive formulas to estimate the standard deviations of the sampling distributions of the estimators.
     a. Thus, we can get an idea of the variability of our estimates.
  3. The shapes of the sampling distributions of the estimators are approximately normal.
     a. Thus, we can perform hypothesis tests in relation to our coefficient estimates.

PART 4

Model Interpretation

Interpreting the Coefficient Estimates

- Suppose we have successfully run MLE. The estimated logistic regression model is given by:

$$\ln\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p = X\hat{\beta}$$

- Upon exponentiating both sides of the equation, we get the relationship:

$$\frac{\hat{p}(X)}{1-\hat{p}(X)} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1 X_1} \times \ldots \times e^{\hat{\beta}_p X_p} = e^{X\hat{\beta}}$$

  - The left hand side of the equation represents the fitted odds of success.
  - The right hand side of the equation is transformed from an additive relationship on the log-odds scale to a multiplicative relationship on the odds scale.

Interpreting the Coefficient Estimates

- Interpretation for β_0 on the log odds scale:
  - When all $X_i = 0$, the log odds of success are β_0.
- Interpretation for β_0 on the odds scale:
  - When all $X_i = 0$, the odds of success are $e^{\beta 0}$.

- Interpretation for β_i ($i = 1, 2, \ldots, p$) on the log odds scale:
  - For a one unit increase in X_i, the log odds of success are increased (or decreased) by β_i, holding all other variables constant.
- Interpretation for β_i ($i = 1, 2, \ldots, p$) on the odds scale:
  - For a one unit increase in X_i, the odds of success are multiplied by $e^{\beta i}$, holding all other variables constant.

Interpreting the Coefficient Estimates

- In general, it's easier to discuss the odds instead of the log odds because this scale makes more intuitive sense to the average individual.

- What if we want to talk about probabilities? We can simply transform the odds back into the probability scale:

$$
\begin{aligned}
O &= \frac{p}{1-p} \\
O(1-p) &= p \\
O - Op &= p \\
O &= p + Op \\
O &= p(1+O) \\
p &= \frac{O}{1+O}
\end{aligned}
$$

Prediction

- Once a valid model is chosen and all β values have been numerically estimated, prediction is quite easy -- except there is one small caveat. Recall that we can determine the probability of success by passing our observations through the function:

$$p(X) = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}$$

- Note that this prediction is a probability -- not a class membership of "success" or "failure" in respect to our dependent variable. We still need to decide on how to define "success" or "failure" based on this probability.

Prediction

- In general, we employ the following decision rule, where c is a cutoff probability:

$$\hat{y}_i = \begin{cases} Success\ (1) & \hat{p}_i \geq c \\ Failure\ (0) & \hat{p}_i < c \end{cases}$$

- In deciding on a cutoff value, we should take into account:
  - The randomness of our data sample.
  - How likely we are to see a success/failure in the population of interest.
  - The costs of incorrectly predicting a success/failure.

- The approach of selecting c = .5 is most commonly used, especially when success and failure are equally likely in the population of interest and when the cost of an incorrect prediction is the same for both success and failure.

PART 5

Assessing Model Fit

Performing Hypothesis Tests: The Wald Test

- A test based on the approximate normality of MLEs is referred to as the Wald test; the Wald test is analogous to the t-tests we used in linear regression.

- For logistic regression, the properties of MLE imply that each coefficient estimate β_i has a sampling distribution that is approximately normal. We can conduct tests of the coefficient estimates by calculating the associated z-test statistic:

$$z^* = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \sim N(0, 1)$$

- The standard error is the estimated standard deviation of the sampling distribution of the coefficient estimate, and will be calculated for us by software.

Performing Hypothesis Tests: The Wald Test

- For logistic regression, the principal hypothesis tests take the following form:
  - Null Hypothesis (H_0): $\beta_i = 0$
  - Alternative Hypothesis (H_A): $\beta_i \neq 0$

- What would it mean if the null hypothesis were true?
  - The log odds of success are unaffected by the value of X_i.
  - In other words, knowing the value of X_i has no bearing on our prediction of success.

- What would it mean if the null hypothesis were false?
  - The log odds of success are affected by the value of X_i.
  - In other words, knowing the value of X_i changes our prediction of success.

Constructing Confidence Intervals

- The construction of confidence intervals is nearly identical in logistic regression as compared to the method we have seen in the past.

- For the coefficients in logistic regression, we can construct an approximate 95% confidence interval as follows:

$$\hat{\beta}_i \pm 2 \cdot SE(\hat{\beta}_i)$$

- Recall that the range of our confidence interval, and thus our uncertainty of the true parameter value, gets larger as the estimate of the standard error gets larger.

Constructing Confidence Intervals

- While we can construct confidence intervals for coefficient estimates, it is often helpful to also construct confidence intervals for the odds ratio associated with each coefficient estimate.

- To construct a 95% confidence interval for the odds ratio for a particular estimate, simply exponentiate the endpoints of the 95% confidence interval for the parameter estimate:

$$e^{\left(\hat{\beta}_i \pm 2 \cdot SE(\hat{\beta}_i)\right)}$$

The Deviance G^2 & Goodness of Fit

- In logistic regression, we replace the concept of the residual sum of squares with the concept of deviance G^2.
  - The deviance, in part, is analogous to the idea of the coefficient of determination used in linear regression, and allows us to compare across various models.

- The deviance G^2 is given by the following equation:

$$G^2 = 2[\ln(l_S) - \ln(l_M)] \sim \chi^2_{n-p}$$

  - The deviance associated with a given logistic regression model M is based on comparing the maximum log-likelihood of model M against the saturated model S.
  - S represents a model that has a parameter fit per observation in our dataset; i.e., this model severely overfits the data at hand.

The Deviance G^2 & Goodness of Fit

- Alongside the deviance, we use a specific application of the X^2 test that evaluates the goodness of fit of the model.
  - Essentially, this X^2 test assesses the fit of the observed values Y as compared to the predicted (expected) values in respect to the model at hand. How well does the model at hand predict the true values?

- The larger the difference between the saturated model and the model at hand (i.e, the larger the deviance), the poorer the fit of the model; if the deviance is large, the estimated model M doesn't really give much of an advantage over the saturated model S.
  - Thus, we want as small a deviance as possible.

The Deviance G^2 & Goodness of Fit

- In logistic regression, the goodness of fit test for the deviance boils down to the following hypotheses:
  - Null Hypothesis (H_0): The logistic regression model M is appropriate.
  - Alternative Hypothesis (H_A): The logistic regression model M is not appropriate.

- **NB:** This overall goodness of fit test is reflective of the overall F-test we saw in multiple linear regression, but the hypotheses seem switched -- be careful!

Using Deviance to Compare Models

- In a similar manner to the use of partial F tests for multiple linear regression, differences in deviances can be used to compare nested models in the logistic regression setting.

- Suppose we have a full logistic regression model with p predictors, and a reduced logistic regression model with a subset of those p predictors. To compare these models against each other, we calculate the drop in deviance:

$$\left(G^2_{Reduced} - G^2_{Full}\right) \sim \chi^2_{df_{Reduced} - df_{Full}}$$

Using Deviance to Compare Models

- The test statistic calculated from the difference between the deviance of two nested logistic regression models performs the following hypothesis test:
  - Null Hypothesis (H_0): The reduced model is sufficient.
  - Alternative Hypothesis (H_A): The reduced model is insufficient.

- This test is directly assessing the difference between G^2_{Reduced} and G^2_{Full}:
  - If this difference is small, then the full model doesn't add much information to the model (H_0).
  - If this difference is large, then the full model does add information to the model (H_A).

Notes on Deviance

- Some common terms surrounding deviance in respect to model building:
  - Null Deviance is calculated based on comparing the saturated model S against the model that only includes a β_0 term (with no predictors); it is a way of assessing the overall maximum deviance because it compares the most complicated model to the most simplistic model.
  - Residual Deviance is calculated based on comparing the saturated model S against the model at hand M; it is a measure of how well M fits the data in general.

- **NB:** The significance of a single term can be tested using the drop in deviance test; however, this is not the same as the Wald test as you would expect from what we saw in multiple linear regression.
  - Should these tests yield differing conclusions, the drop in deviance test produces a more reliable p-value.

McFadden's Pseudo R^2_{dev}

- Recall that for linear regression that R^2 measured the amount of variability in our response that was accounted for by our predictor variables:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Because the deviance G^2 is a generalization of the RSS from linear regression, we can construct a similar measure for the logistic regression setting called McFadden's Pseudo R^2 or R^2_{dev}:

$$R^2_{dev} = 1 - \frac{G^2_M}{G^2_{Null}}$$

- As before, R^2_{dev} is bound between 0 and 1; a higher value indicates a stronger fit.

What About Assumptions & Diagnostics?

- One main difference between linear regression and logistic regression is that in the latter scenario, scatterplots and residual plots tend to be of little value.
  - In logistic regression, only two distinct values of the response variable are possible, so scatterplots often don't yield much information.

- In logistic regression, why is there no burden to check for:
  - Linearity?
  - Constant Variance?
  - Normality?

- However, we still care about the assumptions of independent errors and multicollinearity.

Variable/Model Selection

- Analysis proceeds in a similar fashion as in multiple linear regression analysis; however, since it is difficult to assess model fit from scatterplots or residual analysis, model comparisons comprise most of the exploration.

- Along with the added measure of deviance, the ideas of variable and model selection procedures carry over:
  - Measures of AIC and BIC can be used to compare across models.
  - Forward, backward, and both stepwise procedures can be used to find a suitable model from a subset of the 2^p possibilities.

PART 6

Review

Review

- Part 1: Generalized Linear Models
  - What are Generalized Linear Models?
  - GLMs: Mathematically
- Part 2: Logistic Regression
  - Recall: Simple Linear Regression
  - Binary Response Variables
    - The Problem
    - The Solution
  - Probability & Odds
  - Log Odds as the Link Function
- Part 3: Maximum Likelihood Estimation
  - Estimating the Coefficients
  - Notes on MLE Calculation
  - Properties of MLE
- Part 4: Model Interpretation
  - Interpreting the Coefficient Estimates
  - Prediction
- Part 5: Assessing Model Fit
  - Performing Hypothesis Tests: The Wald Test
  - Constructing Confidence Intervals
  - The Deviance G^2 & Goodness of Fit
  - Using Deviance to Compare Models
  - Notes on Deviance
  - McFadden's Pseudo R^2_{dev}
  - What About Assumptions & Diagnostics?
  - Variable/Model Selection