Logistic Regression Inference

Brooke Coneeny

# Logistic Model Assumptions

## Assumption 1: Appropriate Outcome Type

Logistic regression is an appropriate model to use when the dependent variable is a binary output (1 or 0) which is equivalent to ‘yes’ or ‘no’. If the dependent variable has more than 2 outcomes, a multinomial or ordinal logistic regression is more appropriate.

### Check

We can check for this assumption in our dataset by calculating the number of unique outcomes

## Assumption 2: Linearity of Indp var. & log-odds

An important assumption for logistic regression is that the relationship between the logit (log-odds) of the outcome and each continuous independent variable is linear.

The logit is the logarithm of the odds ratio with p being the probability of a positive outcome (1)

logit() =

### Check

1. Box-Tidwell Test:

* This test checks for the linearity between the predictors and the logit by adding log-transformed interaction terms between the continuous independent variables and their corresponding natural log into the model.
* We then check the statistical significance of the interaction terms based on their p-values.
* If an interaction term has a p-value above a chosen alpha, it implies that the independent variable is linearly related to the logit of the outcome variable and the assumption is satisfied.
* If an interaction term has a p-value below a chosen alpha, meaning it is significant, than there is evidence of non-linearity between the variable and the logit.

1. Visual Check:

* We can also check logit linearity visually by creating a scatter plot between each predictor and the logit values.

## Assumption 3: No Strongly Influential Outliers

Logistic regression assumes that there are no very influential outlier data points. The presence of these points can change the outcome and accuracy of your model.

### Check

1. The influence of points can be checked by using their Cook’s Distance. The Cook’s Distance uses the residual and leverage value of a point to summarize the changes in the regression model when that particular observation is removed. – There are many different cut-off values that are used when deciding which points to classify as influential. A standard threshold is 4/N, where N = number of observations. This means observations with Cook’s Distnace > 4/N are considered to be influential.
2. To determine if a point is an outlier, it standard practice to use the standardized residuals. Data points with an absolute standardized residual value greater than 3 is possibly an extreme outlier

## Assumption 4: Absence of Multicollinearity

Multicollinearity is when the data contains highly correlated independent variables. This can become problematic because it reduces the accuracy of the estimated coefficients which then weakens the models statistical power

### Check

The Variance Inflation Factor (VIF) measures the degree of multicollinearity in a dataset. The VIF is equal to the ratio of the overall model variance to the variance of a model that includes only that sinlge independent variable. - The smallest possible VIF value is 1, meaning there is a complete absence of collinearity between variables - A common rule of thumb, a VIF greater than 5 indicates problematic amount of multicollinearity

## Assumption 5: Independence of Observations

The observations in the dataset must all be independent of each other; meaning they should not come from repeated or paired data. This ensures that each observation is not influenced or related to the others.

### Assumption 6: Sufficientyl Large Sample Size

In order to avoid overfitting the model, there should be a sufficiently large amount of observations for each independent variable

# Fitting Models

We are going to fit two simple logistic regression models. The first will fit y = Home.win on Vegas.H.R

##   
## Call:  
## glm(formula = response1 ~ Vegas.H.R, data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8857 -0.4350 0.2270 0.4136 0.7340   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.491364 0.032428 15.152 < 2e-16 \*\*\*  
## Vegas.H.R 0.028168 0.005493 5.128 5.82e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.2251973)  
##   
## Null deviance: 63.121 on 255 degrees of freedom  
## Residual deviance: 57.200 on 254 degrees of freedom  
## AIC: 348.85  
##   
## Number of Fisher Scoring iterations: 2

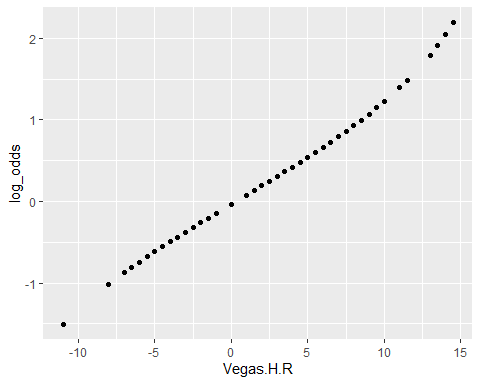
The second regression will fit y = Home.win on Halftime.H.R

##   
## Call:  
## glm(formula = response1 ~ Halftime.H.R, data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.07471 -0.36816 0.02623 0.35427 0.88419   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.544794 0.025857 21.07 <2e-16 \*\*\*  
## Halftime.H.R 0.025234 0.002346 10.76 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.17074)  
##   
## Null deviance: 63.121 on 255 degrees of freedom  
## Residual deviance: 43.368 on 254 degrees of freedom  
## AIC: 277.98  
##   
## Number of Fisher Scoring iterations: 2

Let us quickly go through the assumptions:

1. The outcome is appropriate, there are two choices
2. Linearity of independent variables and log-odds: plot passes for model 1 but not for model 2

# get predicted values and log odds  
modeldata <- modeldata %>%  
 mutate(predicted\_values = predict(model1)) %>%  
 mutate(log\_odds = log(predicted\_values/(1- predicted\_values)))   
  
assumption\_plot <- modeldata %>%  
 ggplot(aes(x = Vegas.H.R, y = log\_odds)) +  
 geom\_point()   
  
assumption\_plot

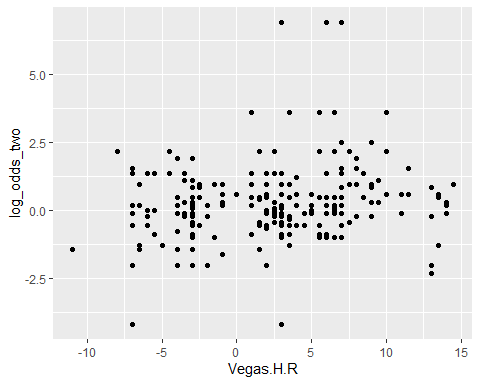


# get predicted values and log odds  
modeldata <- modeldata %>%  
 mutate(predicted\_values2 = predict(model2)) %>%  
 mutate(log\_odds\_two = log(predicted\_values2/(1- predicted\_values2)))

## Warning in log(predicted\_values2/(1 - predicted\_values2)): NaNs produced

assumption\_plot2 <- modeldata %>%  
 ggplot(aes(x = Vegas.H.R, y = log\_odds\_two)) +  
 geom\_point()   
  
assumption\_plot2

## Warning: Removed 23 rows containing missing values (geom\_point).



Explain how to compute the test statistic

Explain how to compute the degrees of freedom

Explain how to compute the lack of fit tests

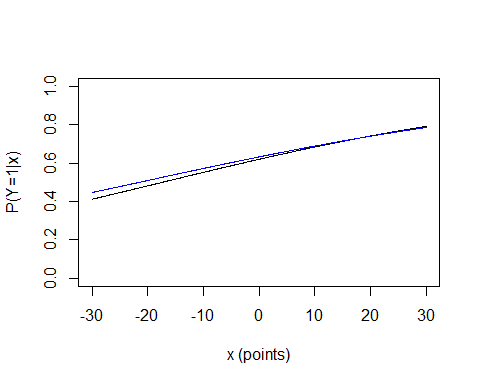
Explain the results and what they imply

# Model Formula

The logistic model formula is as follows: P( | X) =

As for our specific examples from above: model 1: P( | X) = model 2: P( | X) =

# make an empty graph with the correct limits   
plot(0,0, xlim = c(-30,30), ylim = c(0,1), type = "n", xlab = "x (points)", ylab = "P(Y=1|x)")  
  
# add the two lines (change beta0 and beta1 to numbers from your fits)   
# model 1  
lines(seq(-30,30,.1), 1/(1+exp(-(0.49136+0.02817\*seq(-30,30,.1)))))   
# model 2   
lines(seq(-30,30,.1), 1/(1+exp(-(0.54479+0.02523\*seq(-30,30,.1)))), col="blue")



# Hypothesis Testing

#### 

Carry out approximate tests of Ho: (based on standard errors) for model 1

#### 

#### 

Carry out approximate tests of Ho: (based on standard errors) for model 2

#### 

#### 

Explain what the null and alternative imply for model 1

#### 

#### 

Explain what the null and alternative imply for model 2

#### 

# Multiple Logisitic Regression

The above examples only include one explanatory variable, but now let us walk through an example with multiple variables.

model3 <- glm(response1 ~ Vegas.H.R + Halftime.H.R, data = modeldata)  
summary(model3)

##   
## Call:  
## glm(formula = response1 ~ Vegas.H.R + Halftime.H.R, data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.02705 -0.37206 0.01958 0.34411 0.91601   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.509128 0.027811 18.306 < 2e-16 \*\*\*  
## Vegas.H.R 0.015411 0.004882 3.157 0.00179 \*\*   
## Halftime.H.R 0.023195 0.002394 9.687 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1649198)  
##   
## Null deviance: 63.121 on 255 degrees of freedom  
## Residual deviance: 41.725 on 253 degrees of freedom  
## AIC: 270.09  
##   
## Number of Fisher Scoring iterations: 2

This results in the following model: = 0.50913 + 0.01541(Vegas.H.R) + 0.02320(Halftime.H.R)

Therefore, if we wanted to know the fitted probability of the home team winning in the instance that the team is favored by 7 to win and leading by 3 at halftime (Vegas.H.R = 7 and Halftime.H.R = 3), we simply plug in the values 7 and 3 to our equation above:

= 0.50913 + 0.01541(7) + 0.02320(3) = 0.6866

### Explain the interpretations of the fitted coeffificnets

The interpretation for beta coefficients for multiple logistic regression is slightly different than normal linear regression. Let us interpret the above example:

* : When the home team is not favored to win at all (Vega.H.R = 0) and they are not leading at halftime, the log odds of them winning is 0.509
* : For every one unit increase in Vega.H.R , the log odds of the home team winning increases by 0.015
* : For every one unit increase in Halftime.H.R , the log odds of the home team winning increases by 0.023

#### 

Explain the interpretations for the three approximate z test p values ####

#### 

Describe the alternative approximate GLR tests ####

# Sources

<https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290>