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, then 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.

1. 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, family = "binomial", data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9328 -1.0601 0.7242 1.0195 1.6323   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.03457 0.14223 -0.243 0.808   
## Vegas.H.R 0.12390 0.02626 4.718 2.38e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 351.37 on 255 degrees of freedom  
## Residual deviance: 326.36 on 254 degrees of freedom  
## AIC: 330.36  
##   
## Number of Fisher Scoring iterations: 4

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

##   
## Call:  
## glm(formula = response1 ~ Halftime.H.R, family = "binomial",   
## data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7455 -0.8333 0.2253 0.8129 2.2836   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.27694 0.15563 1.779 0.0752 .   
## Halftime.H.R 0.16517 0.02244 7.359 1.85e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 351.37 on 255 degrees of freedom  
## Residual deviance: 251.48 on 254 degrees of freedom  
## AIC: 255.48  
##   
## Number of Fisher Scoring iterations: 5

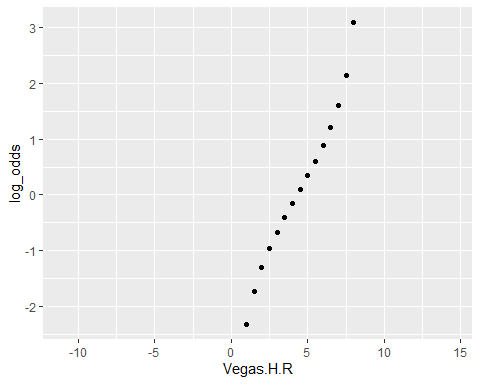
I wanted to quickly show how to check the second asusmption of linearity in case it was not clear above: (2) Linearity of independent variables and log-odds

# get predicted values and log odds  
modeldata <- modeldata %>%  
 mutate(predicted\_values = predict(model1)) %>%  
 mutate(log\_odds = log(predicted\_values/(1- predicted\_values)))

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

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

## Warning: Removed 119 rows containing missing values (geom\_point).

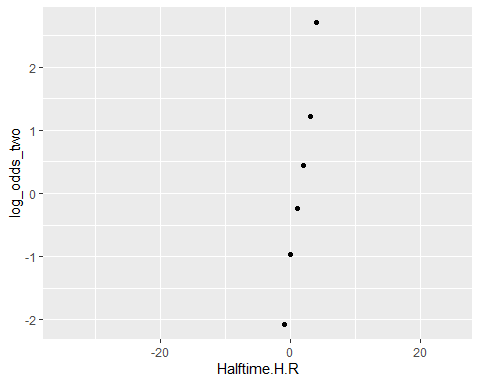


# get predicted values and log odds  
modeldata <- modeldata %>%  
 mutate(predicted\_values2 = predict(model2)) %>%  
 mutate(log\_odds\_two = log(predicted\_values2/(1- predicted\_values2)))

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

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

## Warning: Removed 190 rows containing missing values (geom\_point).



##   
## Call:  
## glm(formula = response1 ~ factor(Vegas.H.R), family = "binomial",   
## data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.03933 -1.08424 0.00022 1.01077 1.87502   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.757e+01 3.956e+03 -0.004 0.996  
## factor(Vegas.H.R)-8 3.513e+01 5.595e+03 0.006 0.995  
## factor(Vegas.H.R)-7 1.757e+01 3.956e+03 0.004 0.996  
## factor(Vegas.H.R)-6.5 1.647e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)-6 1.716e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)-5.5 1.647e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)-5 -2.586e-07 5.595e+03 0.000 1.000  
## factor(Vegas.H.R)-4.5 3.513e+01 4.845e+03 0.007 0.994  
## factor(Vegas.H.R)-4 1.631e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)-3.5 1.797e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)-3 1.600e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)-2.5 3.513e+01 4.423e+03 0.008 0.994  
## factor(Vegas.H.R)-2 1.797e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)-1.5 -2.588e-07 4.845e+03 0.000 1.000  
## factor(Vegas.H.R)-1 1.757e+01 3.956e+03 0.004 0.996  
## factor(Vegas.H.R)0 3.513e+01 5.595e+03 0.006 0.995  
## factor(Vegas.H.R)1 1.951e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)1.5 1.757e+01 3.956e+03 0.004 0.996  
## factor(Vegas.H.R)2 1.757e+01 3.956e+03 0.004 0.996  
## factor(Vegas.H.R)2.5 1.734e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)3 1.757e+01 3.956e+03 0.004 0.996  
## factor(Vegas.H.R)3.5 1.895e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)4 1.785e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)4.5 1.866e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)5 1.647e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)5.5 1.734e+01 3.956e+03 0.004 0.997  
## factor(Vegas.H.R)6 1.841e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)6.5 1.797e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)7 1.858e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)7.5 1.866e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)8 3.513e+01 4.334e+03 0.008 0.994  
## factor(Vegas.H.R)8.5 1.866e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)9 3.513e+01 4.334e+03 0.008 0.994  
## factor(Vegas.H.R)9.5 3.513e+01 4.568e+03 0.008 0.994  
## factor(Vegas.H.R)10 3.513e+01 4.568e+03 0.008 0.994  
## factor(Vegas.H.R)11 3.513e+01 4.845e+03 0.007 0.994  
## factor(Vegas.H.R)11.5 3.513e+01 4.845e+03 0.007 0.994  
## factor(Vegas.H.R)13 1.757e+01 3.956e+03 0.004 0.996  
## factor(Vegas.H.R)13.5 1.826e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)14 1.895e+01 3.956e+03 0.005 0.996  
## factor(Vegas.H.R)14.5 3.513e+01 5.595e+03 0.006 0.995  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 351.37 on 255 degrees of freedom  
## Residual deviance: 268.78 on 215 degrees of freedom  
## AIC: 350.78  
##   
## Number of Fisher Scoring iterations: 16

##   
## Call:  
## glm(formula = response1 ~ factor(Halftime.H.R), family = "binomial",   
## data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.14597 -0.73248 0.00013 0.75853 1.89302   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.857e+01 6.523e+03 -0.003 0.998  
## factor(Halftime.H.R)-33 4.632e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-31 4.632e-07 7.989e+03 0.000 1.000  
## factor(Halftime.H.R)-28 4.638e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-22 4.633e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-21 4.629e-07 7.989e+03 0.000 1.000  
## factor(Halftime.H.R)-18 4.630e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-17 1.696e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-15 4.633e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-14 4.632e-07 6.973e+03 0.000 1.000  
## factor(Halftime.H.R)-13 4.633e-07 7.293e+03 0.000 1.000  
## factor(Halftime.H.R)-11 1.806e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-10 1.696e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-9 4.633e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-8 4.635e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)-7 1.739e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-6 1.816e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-5 3.713e+01 9.224e+03 0.004 0.997  
## factor(Halftime.H.R)-4 1.857e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-3 1.857e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-2 1.926e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)-1 1.816e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)0 1.816e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)1 1.885e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)2 4.633e-07 9.224e+03 0.000 1.000  
## factor(Halftime.H.R)3 2.003e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)4 1.966e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)6 1.995e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)7 1.958e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)8 3.713e+01 7.532e+03 0.005 0.996  
## factor(Halftime.H.R)9 3.713e+01 9.224e+03 0.004 0.997  
## factor(Halftime.H.R)10 3.713e+01 6.752e+03 0.005 0.996  
## factor(Halftime.H.R)11 3.713e+01 7.293e+03 0.005 0.996  
## factor(Halftime.H.R)13 1.926e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)14 3.713e+01 6.918e+03 0.005 0.996  
## factor(Halftime.H.R)15 3.713e+01 7.989e+03 0.005 0.996  
## factor(Halftime.H.R)17 1.995e+01 6.523e+03 0.003 0.998  
## factor(Halftime.H.R)18 3.713e+01 7.532e+03 0.005 0.996  
## factor(Halftime.H.R)20 3.713e+01 9.224e+03 0.004 0.997  
## factor(Halftime.H.R)21 2.076e+01 6.523e+03 0.003 0.997  
## factor(Halftime.H.R)23 3.713e+01 9.224e+03 0.004 0.997  
## factor(Halftime.H.R)24 3.713e+01 7.293e+03 0.005 0.996  
## factor(Halftime.H.R)25 3.713e+01 9.224e+03 0.004 0.997  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 351.37 on 255 degrees of freedom  
## Residual deviance: 220.69 on 213 degrees of freedom  
## AIC: 306.69  
##   
## Number of Fisher Scoring iterations: 17

# Computing the test statistic

For model 1:

To compare our logistic model to our saturated model we are going to use the GLR test. The GLR test finds the difference between the max likelihood for the logistic model (which has two betas) and the saturated log-likelihood model (which has 41 betas).

max likelihood for logistic = -163.18141

max likelihood for log-like = -134.3914

The difference = 28.79

Once we have this difference, we multiply it by negative 2 to get 57.58

This statistic acts as a chi-squared distribution with degrees of freedom df = 41 - 2 = 39

The test statistic has a p-value of 0.0279, telling us it is significant at the alpha value of 0.05. Because it is significant this suggests that the logistic model may be incorrect

For model 2:

We are now going to do the same process except for the second model, which uses the half time score margin to predict the winner of the game.

max likelihood for logistic = -125.74107

max likelihood for log-like = -110.34566

The difference = 15.39541

We then multiply this difference by negative 2 to get -30.79082

This statistic acts as a chi-squared distribution with degrees of freedom df = 41 - 2 = 39

The test statistic has a p-value of 0.8774 which is not statistically significant at the alpha = 0.05 level or even at the alpha = 0.1 level. This suggests that the logistic model is correct

# Model Formula

The logistic model formula is as follows:

P( | X) =

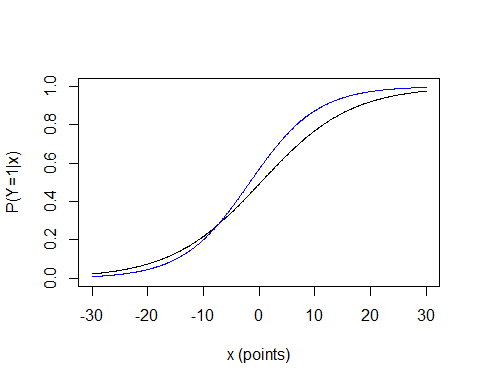
As for our specific examples from above:

model 1: P( | X) =

model 2: P( | X) =

The plot below shows the probability curves of both models. The model which is predicted using the vegas odds is in black and the model which uses the halftime score margin is in blue. In the beginning the two are relatively consistent, but as we can see the using the halftime score allows us to better predict whether the home team wins.

# 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(-(-0.03457+0.12389\*seq(-30,30,.1)))))   
# model 2: uses half time  
lines(seq(-30,30,.1), 1/(1+exp(-(0.27694+0.16517\*seq(-30,30,.1)))), col="blue")



# Hypothesis Testing

### Model 1

First we are going to carry out approximate tests of Ho: (based on standard errors) for model 1. The estimates we are using become approximately normal because we are using the Maximum Likelihood Estimates. This means we can use an approximate z test for the model parameters.

Looking at the parameter estimate output for model 1, we can see the estimate for and the standard error is 0.14223

Our approximate z-statistic = which has a p-value of 0.808. This means the coefficient on for the first model is not significant and we cannot reject the null hypothesis which says is equal to 0.

### Model 2

Now we are going to carry out approximate tests of Ho: (based on standard errors) for model 2. The estimates we are using become approximately normal because we are using the Maximum Likelihood Estimates. This means we can use an approximate z test for the model parameters.

Looking at the parameter estimate output for model 1, we can see the estimate for and the standard error is 0.1556258

Our approximate z-statistic = which has a p-value of 0.0752. This means the coefficient on for the first model is significant at the alpha = 0.1 level. At this level we can reject the null hypothesis which says is equal to 0.

# Multiple Logisitic Regression

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

model5 <- glm(response1 ~ Vegas.H.R + Halftime.H.R, data = modeldata, family = "binomial")  
summary(model5)

##   
## Call:  
## glm(formula = response1 ~ Vegas.H.R + Halftime.H.R, family = "binomial",   
## data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7144 -0.8315 0.2094 0.7613 2.4335   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.05103 0.17215 0.296 0.76690   
## Vegas.H.R 0.10072 0.03174 3.173 0.00151 \*\*   
## Halftime.H.R 0.16219 0.02308 7.027 2.11e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 351.37 on 255 degrees of freedom  
## Residual deviance: 240.76 on 253 degrees of freedom  
## AIC: 246.76  
##   
## Number of Fisher Scoring iterations: 5

This results in the following model:

P( | X) =

P( | X) =

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:

P( | Vegas.H.R = 7 & Halftime.H.R = 3) = = 0.2239768

This means the log odds of the home team winning, when Vegas.H.R = 7 and Halftime.H.R = 3, is 0.2239768

### Explain the interpretations and p-values 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 (Vegas.H.R = 0) and they are not leading at halftime, the log odds of them winning is 0.05103

The p-value for this beta is not significant, which means we cannot reject the null hypothesis that it is 0.

* : For every one unit increase in Vega.H.R , holding constant Halftime.H.R ,the log odds of the home team winning increases by 0.10072
* : For every one unit increase in Halftime.H.R , holding constant Vegas.H.R , log odds of the home team winning increases by 0.16219

The p-value for both and is less than alpha = 0.05, which means they are statistically significant. This means we can reject their individual null hypotheses that they should be equal to 0.

### Alternative approximate GLR tests

We are now going to perform a GLR test comparing our multiple logit model and the model which only uses Halftime.H.R as a predictor.

max likelihood for multiple logit = -120.38

max likelihood for Halftime.H.R model = -125.74

The difference, multiplied by negative 2 gives us: 10.7187431

This statistic acts as a chi-squared distribution with degrees of freedom 1, for the difference in number of parameters between the two models is only 1

The test statistic has a p-value of 0.0011 which is statistically significant at the alpha = 0.05 level. This suggests we can reject the null hypothesis that the beta for Vegas.H.R should be 0. This also tells us that our multiple logit regression is a better estimate than the model which only uses halftime as a predictor.

# Sources

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