

## Logistic Regression in R

Meredith Dodd <meredith.dodd@woz-u.com>

Sun 10/4/2020 4:31 PM

To: Meredith Dodd <meredith.dodd@woz-u.com>

# Predict the helpfulness of the Amazon music reviews using Logistic Regression. Outcome is 0 (not helpful) or 1 (helpful)

## Load in Libraries

```
library("caret")  
library("magrittr")  
library("dplyr")  
library("tidyr")  
library("lmtest")  
library("popbio")  
library("e1071")  
library("IDPmisc")
```

## Data Wrangling

### Recode outcome (DV) to zeros and ones

```
Musical_instruments_reviews$HelpfulYN <- NA  
Musical_instruments_reviews$HelpfulYN[Musical_instruments_reviews$helpful == '[0, 0]'] <- 0  
Musical_instruments_reviews$HelpfulYN[Musical_instruments_reviews$helpful == '[1, 1]'] <- 1
```

### Remove Missing Data

```
Reviews <- NaRV.omit(Musical_instruments_reviews)
```

## Testing Assumptions

### Sample size

#### Run the Base Model

```
mylogit <- glm(HelpfulYN ~ overall, data=Reviews, family="binomial")
```

#### Predict Helpfulness

```
probabilities <- predict(mylogit, type="response")  
Reviews$Predicted <- ifelse(probabilities > .5, "pos", "neg")
```

```
Reviews$PredictedR <- NA  
Reviews$PredictedR[Reviews$Predicted == 'pos'] <- 1  
Reviews$PredictedR[Reviews$Predicted == 'neg'] <- 0
```

```
Reviews$PredictedR <- as.factor(Reviews$PredictedR)  
Reviews$HelpfulYN <- as.factor(Reviews$HelpfulYN)
```

```
conf_mat <- caret::confusionMatrix(Reviews$PredictedR, Reviews$HelpfulYN)
conf_mat

### Do not meet the assumptions for sample size - need at least 1 per cell and we have two with 0 in the
cell

### Logit Linearity

Reviews1 <- Reviews %>% dplyr:: select_if(is.numeric)

predictors <- colnames(Reviews1)

Reviews2 <- Reviews1 %>%
  mutate(logit=log(probabilities/(1-probabilities))) %>%
  gather(key= "predictors", value="predictor.value", -logit)

ggplot(Reviews2, aes(logit, predictor.value)) +
  geom_point(size=.5, alpha=.5) +
  geom_smooth(method="loess") +
  theme_bw() +
  facet_wrap(~predictors, scales="free_y")

### It is roughly linear, so we will move on

### Multicollinearity - only if you have multiple IVs, which we don't (but would test with correlation)

### Independent Errors

plot(mylogit$residuals)

### Looking for an even distribution of points straight across - you mostly have that, but it's concerning
that there is an upper and lower contingent. Let's do a Durbin Watson test to get more info!
dwtest(mylogit, alternative = "two.sided")

### You want this to be > .05, but it isn't - which means that you need to look further at the DW value.
<1 or >3 is a problem...so we are safe and can proceed with testing.

### Screening for Outliers

infl <-influence.measures(mylogit)
summary(infl)

### There are definitely some outliers here, but we will proceed for now

## Examine Output

summary(mylogit)

## The overall score does not seem to be indicative of whether the review was helpful or not
```

## Graph it!

```
logi.hist.plot(Reviews$overall, Reviews$HelpfulYN, boxp=FALSE, type="hist", col="gray")
```

**Meredith Dodd, Ph.D. | Data Science Program Chair and Instructor**

[meredith.dodd@woz-u.com](mailto:meredith.dodd@woz-u.com)

o: 480-291-8068

powered by



<https://woz-u.com>