Biostatistics week 10: logistic regression

Schedule for today

- Peer review (last one)
- Statistics flow chart
- Logistic regression workshop
- Challenge time or work with your book club group

Outline

These slides cover the basics of *logistic regression*:

- Logistic model
- Predictor significance and interpretation
- Model significance
- Model fit
- Assumption checking
- Alternate models

What predicts becoming a zombie?

- Age was significantly associated with zombie status; zombies were older on average
- Rurality was significantly associated with zombie status; more zombies were urban, fewer were rural
- Sex was not associated with zombie status; males and females were equally likely to be zombies
- Having a week of food was associated with zombie status; those without food were more likely to be zombies

Bring in the zombie data

- Data are in a csv file saved online

```
# DATA IMPORT
zombies <- read.csv("http://tinyurl.com/fobzombies")
```

Zombie variables

The zombie data includes the following:

- **zombieid:** unique identifier for each observed person
- **age:** age in years
- **gender:** sex (female/male)
- **rurality:** lives in rural/suburban/urban area
- **household:** how many people live in the household
- **water:** gallons of water available
- **radio:** have a battery powered radio (yes/no)
- **flashlight:** have a flashlight (yes/no)
- **firstaid:** have a first aid kit (yes/no)
- **zombie:** zombie status (zombie, not zombie)

Logistic regression

- Logistic regression follows the principles of linear regression
- The outcome or dependent variable is *binary* (e.g., zombie status)
- The linear regression model is transformed using the *logit transformation* to predict the probability of the outcome of interest

$$p(y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2)}}$$

  - *y is the binary outcome variable*
  - x_1, x_2, *etc are predictors of the outcome (e.g., age, rurality)*
  - *p(y) is the probability of the outcome (e.g., becoming a zombie)*
  - b_0 *is the y-intercept*
  - b_1, b_2, *etc are the slopes for* x_1 x_2

Modeling binary outcomes

- What predicts being a zombie (vs. non-zombie)?
  - *Are the predictors statistically significantly related to zombie status?*
  - *Is the model statistically significantly better than the baseline?*
  - *How well does the model fit?*

Null and alternate hypotheses for the predictors

- H0: There is no relationship between age and zombie status
- HA: There is a relationship between age and zombie status
- H0: There is no relationship between sex and zombie status
- HA: There is a relationship between sex and zombie status
- H0: There is no relationship between rurality and zombie status
- HA: There is a relationship between rurality and zombie status

You try it! Write the null and alternate for food access:

- H0:
- HA:

Estimating a logistic regression

$$p(zombie) = \frac{1}{1 + e^{-(b_0 + b_1 age + b_2 gender + b_3 rurality + b_4 food)}}$$

```
# estimate zombie model with age, sex, rurality
zmodel <- glm(zombie ~ age + gender + rurality + food, data = zombies,
              family = binomial(logit))
summary(zmodel)
```


R Output for model

```
##
## Call:
## glm(formula = zombie ~ age + gender + rurality + food, family = binomial(logit),
##     data = zombies, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2197  -0.6403  -0.2255   0.4716   2.3845
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.07426    0.68868  -4.464 8.04e-06 ***
## age               0.05508    0.01251   4.402 1.07e-05 ***
## gendermale        0.43559    0.40663   1.071  0.28408
## ruralitysuburban  1.53607    0.48114   3.193  0.00141 **
## ruralityurban     2.69970    0.52949   5.099 3.42e-07 ***
## foodyes          -2.56786    0.43544  -5.897 3.70e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.37  on 199  degrees of freedom
## Residual deviance: 157.20  on 194  degrees of freedom
## AIC: 169.2
##
## Number of Fisher Scoring iterations: 5
```


Significance of predictors

- Because the model has been transformed, the "slopes" or coefficients for the predictors cannot be directly interpreted
- Instead, calculate odds ratios and confidence intervals for each predictor:

```
# get odds ratios by exponentiating the coefficients
oddsRatios <- exp(cbind(OR = coef(zmodel), confint(zmodel)))
oddsRatios
```

```
##                          OR      2.5 %     97.5 %
## (Intercept)       0.04622403 0.01112913  0.1682178
## age               1.05662719 1.03203509  1.0842272
## gendermale        1.54586748 0.70094035  3.4840464
## ruralitysuburban  4.64629062 1.84368879 12.2946319
## ruralityurban    14.87531118 5.53667716 44.8149219
## foodyes           0.07669971 0.03099854  0.1729051
```

Odds ratios and statistical significance

- An odds ratio quantifies the increase or decrease in the odds of the outcome depending on the value of the predictor
- An odds ratio of 1 indicates that the odds of the outcome are 1 times higher/lower for one group than for another, so odds ratio of 1 is no difference!
- In the population, is the odds ratio different from 1?
  - *Confidence intervals with 1 in them show that the odds could be 1 in the population*
  - *Confidence intervals without 1 in them show that the odds are signficantly different from 1*
- An odds ratio with a **95% CI that DOES NOT include 1** is **statistically significantly** different from 1…that is, there is some relationship

Interpreting the continuous predictor odds ratios

```
##                         OR      2.5 %     97.5 %
## (Intercept)       0.04622403 0.01112913  0.1682178
## age               1.05662719 1.03203509  1.0842272
## gendermale        1.54586748 0.70094035  3.4840464
## ruralitysuburban  4.64629062 1.84368879 12.2946319
## ruralityurban    14.87531118 5.53667716 44.8149219
## foodyes           0.07669971 0.03099854  0.1729051
```

- For continuous predictors, odds ratios show increase in odds of the outcome for a one-unit increase in the predictor
  - *Reject the null hypothesis that age is not associate with zombie status. There is a statistically significant relationship between age and zombie status. For every one year increase in age, the odds of being a zombie increase 1.06 times (95% CI: 1.03 - 1.08).*

Interpreting the categorical predictor odds ratios

```
##                         OR      2.5 %     97.5 %
## (Intercept)       0.04622403 0.01112913  0.1682178
## age               1.05662719 1.03203509  1.0842272
## gendermale        1.54586748 0.70094035  3.4840464
## ruralitysuburban  4.64629062 1.84368879 12.2946319
## ruralityurban    14.87531118 5.53667716 44.8149219
## foodyes           0.07669971 0.03099854  0.1729051
```

- For categorical predictors, odds ratios show the increase or decrease in odds of the outcome for the category shown compared to the *reference* group (category not shown)
  - *Retain the null hypothesis; there is no statistically significant relationship between sex and zombie status.*
  - *Reject the null hypothesis that rurality is not associate with zombie status. There is a statistically significant relationship between rurality and zombie status. The odds of being a zombie are 4.65 times higher for suburban compared to rural areas (95% CI: 1.84 - 12.29). The odds of being a zombie are 14.88 times higher for urban compared to rural (95% CI: 5.54 - 44.81).*

Interpreting odds ratios less than 1

- Occasionally you will have an odds ratio that is below 1 with a confidence interval that does not cross 1, like OR = .75 with 95% CI: .68-.84
- This can be interpreted in two ways, for example if the odds ratio for .75 were for sex and zombie status:
  - *The odds of being a zombie is .75 times as high for males compare to females (95% CI: .68-.84).*
  - *The odds of being a zombie is 25% lower for males compared to females (OR = .75; 95% CI: .68-.84).*

You try it

```
##                         OR      2.5 %     97.5 %
## (Intercept)       0.04622403 0.01112913  0.1682178
## age               1.05662719 1.03203509  1.0842272
## gendermale        1.54586748 0.70094035  3.4840464
## ruralitysuburban  4.64629062 1.84368879 12.2946319
## ruralityurban    14.87531118 5.53667716 44.8149219
## foodyes           0.07669971 0.03099854  0.1729051
```

- Interpret the odds ratio for the food variable:

Is the model statistically significantly better than the baseline?

- The baseline is the percentage of zombies (or whatever the outcome is):

```
# get baseline probability
prop.table(table(zombies$zombie))
```

```
##
## not zombie     zombie
##      0.605      0.395
```

- The probability of .395 or 39.5% zombies is the baseline
- Without knowing anything else, you would be more likely to predict that each person was NOT a zombie (because 60.5% are NOT zombies)
- Predicting everyone is not a zombie would result in the prediction being right 60.5% of the time
- Can the model do better than that?

Null and Alternate Hypotheses

H0: The model is no better than the baseline percentage at predicting zombie status HA: The model is better than the baseline at predicting zombie status

Get the test statistic and p-value

- Unfortunately the statistical test of the model does not come with the model output, so it has to be computed separately
  - *The test statistic for the model is a* χ^2

```
# model chi-squared
zchi <- with(zmodel, null.deviance - deviance)
# degrees of freedom
zdf <- with(zmodel, df.null - df.residual)
# p-value
zp <- with(zmodel, pchisq(zchi, zdf, lower.tail = FALSE))
# altogether
modelsig <- round(c(zchi, zdf, zp), 3)
names(modelsig) <- c("Chi-squared", "d.f.", "p")
modelsig
```

```
## Chi-squared        d.f.           p
##     111.171       5.000       0.000
```

Reject the null hypothesis, the logistic regression model was significantly better than the baseline in predicting zombie status ($\chi^2(5) = 111.17$; $p < .05$).

Model Fit

- The last thing before putting the interpretation together is model fit
- In linear regression the model fit is R^2, which is the percent of variance in the outcome accounted for by the model
- In logistic regression the model fit is the **percent correctly predicted**
  - *Use the model to predict the probability that each person is a zombie*
  - *If the model predicts 50% or more probability that a person is a zombie, they are predicted to be a zombie*
  - *If the probability they are a zombie is below 50%, the person is predicted to not be a zombie*
  - *Compare the predicted zombie status with the observed zombie status to see what proportion the model predicted correctly!*

Computing Model Fit

```
# predict the probability of being a zombie and round to 0 or 1
zombies$zpredict <- round(predict(zmodel, type = "response"))
# add labels of zombe not zombie
zombies$zpredict <- ifelse(zombies$zpredict == 1, "zombie", "not zombie")

# make a table of observed and predicted
library(descr)
CrossTable(zombies$zombie, zombies$zpredict,
           prop.c = FALSE, prop.chisq = FALSE, prop.t = FALSE)
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
## ============================================
##                   zombies$zpredict
## zombies$zombie    not zombie   zombie   Total
## --------------------------------------------
## not zombie               105       16     121
##                        0.868    0.132   0.605
## --------------------------------------------
## zombie                    18       61      79
##                        0.228    0.772   0.395
## --------------------------------------------
## Total                    123       77     200
## ============================================
```

- Columns are predicted, rows are observed

Interpret the model fit

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
## ============================================
##                   zombies$zpredict
## zombies$zombie    not zombie   zombie   Total
## --------------------------------------------
## not zombie               105       16     121
##                        0.868    0.132   0.605
## --------------------------------------------
## zombie                    18       61      79
##                        0.228    0.772   0.395
## --------------------------------------------
## Total                    123       77     200
## ============================================
```

- 166 were correctly predicted by the model out of 200
  - *105 (86.8%) of not zombies were predicted to be not zombies (correct)*
  - *16 (13.2%) of not zombies were predicted to be zombies (incorrect)*
  - *18 (22.8%) of zombies were predicted to be not zombies (incorrect)*
  - *61 (77.2%) of zombies were predicted to be zombies (correct)*

The model correctly predicted 166/200 or 83% of the time. It was better at predicting non-zombies (86.8% correct) compared to zombies (77.2% correct).

Altogether

- Model significance
- Model fit
- Predictor odds ratios and CI

A logistic regression model including age, sex, and rurality as predictors of zombie status was statistically significantly better than the baseline at explaining zombie status ($\chi^2(5) = 111.17$; $p < .05$). The model correctly predicted 83% of observations including 86.8% of the non-zombies and 77.2% of the zombies. The odds of being a zombie was not statistically significantly different for males compared to females. For every one year increase in age, the odds of being a zombie increased 1.06 times (95% CI: 1.03 - 1.08). The odds of being a zombie was 4.65 times higher for people in suburban compared to rural areas (95% CI: 1.84 - 12.29). The odds of being a zombie is 14.88 times higher for people in urban compared to rural areas (95% CI: 5.54 - 44.81). The odds of being a zombie is 93% lower for those with food compared to those without (OR = .07; 95% CI: .03 - .17).

Assumptions

- Independent observations (not tested, just known)
- No multicollinearity
- Linearity of independent variables with the log-odds of the outcome

Testing multicollinearity

- Multicollinearity is when two variables are highly correlated with one another and so are redundant
- It is identified by the VIF in linear regression and the GVIF in logistic
  - *The GVIF, or generalized variance inflation factor, re-runs the model for each predictor as the outcome with the other predictors as the independent variables*
  - *If model fit is very high for these models, that means the predictors are strongly related to each other and do not all need to be in the model*
  - *Bottom line, if the GVIF score is too high, a variable is too strongly related to another variable and one of them should be removed*
  - *A variable should be removed if* $GVIF^{1/(2*df)} > 4$

```
# compute GVIF for zmodel
library(car)
vif(zmodel)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.130236  1        1.063126
## gender   1.032017  1        1.015882
## rurality 1.148817  2        1.035292
## food     1.176003  1        1.084437
```

- The last column shows the $GVIF^{1/(2*df)}$ and none are > 4
- This assumption is MET

Testing linearity with log odds

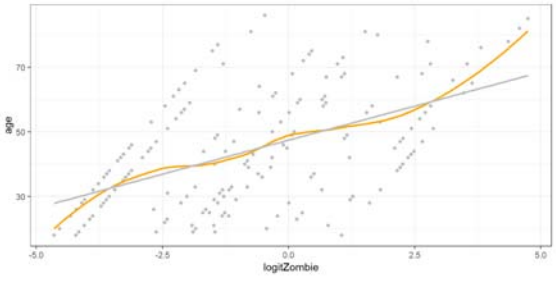
- This assumption is only tested for continuous predictors in the model
  - *age is the only continuous predictor*
- Test by plotting the predictor value on the y-axis and the logit of the outcome on the x-axis
- Minor deviations from linearity are ok, if monotonic

```
# predict the probability of zombie-ness
zombies$zpredictProb <- predict(zmodel, type = "response")
# make a variable of the logit of the outcome
zombies$logitZombie <- log(zombies$zpredictProb/(1-zombies$zpredictProb))

# graph the logit of the outcome variable with the predictor of age
library(ggplot2)
ggplot(zombies, aes(x = logitZombie, y = age))+
  geom_point(color = "gray") +
  geom_smooth(se = FALSE, color = "orange") +
  geom_smooth(method = lm, se = FALSE, color = "gray") +
  theme_bw()
```


Linearity graph

- There are deviations from linearity at both ends, but it generally follows the line and is a monotonic relationship (just goes one direction)
- Assumption is MET



Alternatives for not meeting assumptions

There is no one specific test that is the one alternative to logistic regression when assumptions are not met. Some of the options for dealing with failed assumptions are:

- Include additional variables in the model and check assumptions again
- Recode the independent variable(s) into categories and try again
- Use an alternate model for binary outcomes like negative binomial regression or tweedie regression

The End

- 2 options for (your very last challenge!) Challenge 10, complete ONE of these two:
  - *DataCamp Multiple and Logistic Regression Course Chapter 4 (If you do this, please submit a note in Blackboard that you completed the DataCamp version of the challenge)*
  - *Usual style of challenge in posted in GitHub*