Logistic Regression

Dr. Tackett

11.07.2018
Announcements

- HW 6 due today
- Lab 8 due Sunday, 11/11
- Project Proposal due Tuesday, 11/13
  - Example posters in office
- HW 7 due Thursday, 11/15
Packages

library(knitr)
library(broom)
library(dplyr)
library(tibble)
library(ggplot2)
library(pROC)  #ROC curves
#library(arm)  #binned residuals
library(NHANES)  #subset of NHANES dataset
library(questionr)  #odds ratio function
Review

- $Y$: binary response
  - 1: yes
  - 0: no

- $Mean(Y) = \pi$

- $Var(Y) = \pi(1 - \pi)$

- Odds of "yes": $\omega = \frac{\pi}{1-\pi}$
Comparing Odds

Suppose we have two independent groups with odds $\omega_1$ and $\omega_2$

- **Odds Ratio:** $\phi = \frac{\omega_1}{\omega_2}$

- Use inference to assess if groups have equal odds, i.e. $\phi = 1$
  
  - **Hypothesis Test:**
  
  $$H_0 : \log(\phi) = 0$$

  - **Confidence Interval:**
  
  $$\exp \left\{ \log(\phi) \pm z^{*} SE(\log(\phi)) \right\}$$
NHANES Data

- National Health and Nutrition Examination Survey is conducted by the National Center for Health Statistics (NCHS)

- The goal is to "assess the health and nutritional status of adults and children in the United States"

- This survey includes an interview and a physical examination
NHANES Data

- We will use the data from the NHANES R package

- Contains 75 variables for the 2009 - 2010 and 2011 - 2012 sample years

- The data in this package is modified for educational purposes and should **not** be used for research

- Original data can be obtained from the NCHS website for research purposes

- Type `?NHANES` in console to see list of variables and definitions
NHANES: Physical Activity & Sleep

- Do people who do regular physical activity have lower odds of sleep problems than those who do not do regular physical activity?

- We will analyze the following variables:
  - **PhysActive**: Participant does moderate to vigorous-intensity sports, fitness or recreational activities
  
  - **SleepTrouble**: Participant has told doctor or other health professional they had trouble sleeping
NHANES: Physical Activity & Sleep

# Filter data to only include adults with all data for pulse
set.seed(1234)
nhanes <- NHANES %>% filter(Age > 18, !is.na(Pulse)) %>%
  sample_n(500)
nhanes %>% select(Age, Gender, PhysActive, SleepTrouble) %>% glimpse()

## Observations: 500
## Variables: 4
## $ Age <int> 59, 32, 24, 29, 58, 62, 80, 30, 52, 44, 57, 32, 5...
## $ Gender <fct> male, female, female, male, female, male, male, m...
## $ PhysActive <fct> No, No, No, No, Yes, No, No, Yes, Yes, No, No, No...
## $ SleepTrouble <fct> Yes, No, Yes, No, No, No, No, No, Yes, No, No, No...
#get counts for each combination of PhysActive and SleepTrouble
```r
table <- nhanes %>% group_by(PhysActive, SleepTrouble) %>% summarise(kable(table, format="html")
```

<table>
<thead>
<tr>
<th>PhysActive</th>
<th>SleepTrouble</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>161</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>75</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>189</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>75</td>
</tr>
</tbody>
</table>
NHANES: Physical Activity & Sleep

- Group 1: PhysActive="Yes", Group 2: PhysActive="No"
  - \( \omega_1 : 0.3968 \)
  - \( \omega_2 : 0.4658 \)
  - \( \phi : 0.8519 \)
  - \( SE(\log(\phi)) = 0.1954 \)
  - \( SE_0(\log(\phi)) = 0.1955 \)

- Is there evidence that people who do regular physical activity have lower odds of sleep problems than those who do not do regular physical activity?

- Calculate a 95% confidence interval for the odds ratio.
NHANES: Physical Activity & Sleep

- R has a function to calculate the odds ratio

- `odds.ratio()` function in the `questionr` package is shown below

```r
# calculate odds ratio and 95% confidence interval
model <- glm(SleepTrouble ~ PhysActive, family=binomial, data=nhane)
or <- odds.ratio(model, level=0.95)
kable(or, format="html", digits=3)
```

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>2.5 %</th>
<th>97.5 %</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.466</td>
<td>0.352</td>
<td>0.61</td>
<td>0.000</td>
</tr>
<tr>
<td>PhysActiveYes</td>
<td>0.852</td>
<td>0.580</td>
<td>1.25</td>
<td>0.412</td>
</tr>
</tbody>
</table>
We want to build a model to incorporate more variables that could potentially explain the odds of having sleep problems.
Linear model?

- We want to use a model to predict a binary response $Y$
Linear model?

- We want to use a model to predict a binary response \( Y \)

- Suppose we use a linear regression model to predict \( Y \) using some explanatory variable \( X \)

\[
Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)
\]
Linear model?

- We want to use a model to predict a binary response $Y$

- Suppose we use a linear regression model to predict $Y$ using some explanatory variable $X$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

- This model assumes that $Y$ could be any continuous value; however, it can only be 0 or 1
Linear model?

- We want to use a model to predict a binary response $Y$

- Suppose we use a linear regression model to predict $Y$ using some explanatory variable $X$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

- This model assumes that $Y$ could be any continuous value; however, it can only be 0 or 1

- So linear regression is not appropriate
Other model choices

Let $P(Y_i = 1|X_i) = \pi_i$ and $P(Y_i = 0|X_i) = 1 - \pi_i$
Other model choices

Let $P(Y_i = 1|X_i) = \pi_i$ and $P(Y_i = 0|X_i) = 1 - \pi_i$

Potential models for $\pi_i$: 
Other model choices

Let $P(Y_i = 1|X_i) = \pi_i$ and $P(Y_i = 0|X_i) = 1 - \pi_i$

Potential models for $\pi_i$:

- **Linear**: $\pi_i = \beta_0 + \beta_1 X_i$
  - $\pi_i$ possibly outside of $(0, 1)$
Other model choices

Let $P(Y_i = 1|X_i) = \pi_i$ and $P(Y_i = 0|X_i) = 1 - \pi_i$

Potential models for $\pi_i$:

- **Linear**: $\pi_i = \beta_0 + \beta_1 X_i$
  - $\pi_i$ possibly outside of $(0, 1)$

- **Log-linear**: $\log(\pi_i) = \beta_0 + \beta_1 X_i$
  - $\pi_i$ possibly greater than 1
Logistic Regression Model

- Suppose $P(Y_i = 1|X_i) = \pi_i$ and $P(Y_i = 0|X_i) = 1 - \pi_i$

- The **logistic regression model** is

  $$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i$$

- $\log \left( \frac{\pi_i}{1 - \pi_i} \right)$ is called the **logit** function
Logistic Regression Model

\[ \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i \]

- We can calculate \( \pi_i \) by solving the logit equation:

\[ \pi_i = \frac{\exp{\beta_0 + \beta_1 X_i}}{1 + \exp{\beta_0 + \beta_1 X_i}} \]
Solving Logit Equation

\[ \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i \]

\[ \Rightarrow \quad \exp \left\{ \log \left( \frac{\pi_i}{1 - \pi_i} \right) \right\} = \exp \{\beta_0 + \beta_1 X_i\} \]

\[ \Rightarrow \quad \frac{\pi_i}{1 - \pi_i} = \exp \{\beta_0 + \beta_1 X_i\} \]

\[ \Rightarrow \quad \pi_i = \frac{\exp \{\beta_0 + \beta_1 X_i\}}{1 + \exp \{\beta_0 + \beta_1 X_i\}} \]
Interpreting Model Coefficients

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i$$
Interpreting Model Coefficients

\[
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i
\]

- **Slope, \( \beta_1 \):**
  - As \( X_i \) increases by 1 unit, log-odds of \( Y \) increases by \( \beta_1 \)
  - As \( X_i \) increases by 1 unit, the odds of \( Y \) multiply by a factor of \( \exp \{ \beta_1 \} \)
Interpreting Model Coefficients

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i$$
Interpreting Model Coefficients

\[
\log\left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i
\]

- **Intercept, \(\beta_0\):**
  - When \(X = 0\), log-odds of \(Y\) are \(\beta_0\)
  - When \(X = 0\), odds of \(Y\) are \(\exp\{\beta_0\}\)
  - If we mean-center \(X\), then we can interpret the intercept in terms of the mean of \(X\)
Interpreting Model Coefficients

\[
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i
\]

- **Intercept, \( \beta_0 \):**
  - When \( X = 0 \), log-odds of \( Y \) are \( \beta_0 \)
  - When \( X = 0 \), odds of \( Y \) are \( \exp \{ \beta_0 \} \)
  - If we mean-center \( X \), then we can interpret the intercept in terms of the mean of \( X \)

- Can interpret results by graphing predicted \( \pi \) for values of \( X \)
Estimating the Coefficients

- Estimate coefficients using *maximum likelihood estimation*
  - covered in STA 250 and STA 360
Estimating the Coefficients

- Estimate coefficients using *maximum likelihood estimation*
  - covered in STA 250 and STA 360

- **Basic Idea:**
  - Find values of $\beta_0$ and $\beta_1$ that make observed values of $Y$ the most likely to have occurred
  - Use multivariable calculus and numerical methods to estimate coefficients
Estimating the Coefficients

- Estimate coefficients using *maximum likelihood estimation*
  - covered in STA 250 and STA 360

- **Basic Idea:**
  - Find values of $\beta_0$ and $\beta_1$ that make observed values of $Y$ the most likely to have occurred
  - Use multivariable calculus and numerical methods to estimate coefficients

- We will use R to estimate the coefficients
Inference for Coefficients

- **standard error, $SE(\hat{\beta}_1)$**: estimated standard deviation of the sampling distribution of $\hat{\beta}_1$

- We can calculate the $100(1 - \alpha)\%$ confidence interval based on the large-sample Normal approximations
  
  - CI for $\beta_1$:
    
    $$\hat{\beta}_1 \pm z^* SE(\beta_1)$$

  - CI for $\exp\{\beta_1\}$:
    
    $$\exp\{\hat{\beta}_1 \pm z^* SE(\beta_1)\}$$
Logistic Regression in R

- Use the \texttt{glm()} function in R

- Set \texttt{family=binomial}

\begin{verbatim}
my.model <- glm(Y ~ X, family=binomial, data=my.data)
\end{verbatim}
Can pulse rate be used to distinguish the odds of an adult having trouble sleeping?

**Pulse**: 60 second pulse rate

```
# calculate logistic model using mean-centered Pulse
nhanes <- nhanes %>% mutate(pulse.cent = Pulse - mean(Pulse))
log.odds <- glm(SleepTrouble ~ pulse.cent, family=binomial, data=nhanes)
kable(tidy(log.odds), format="html", digits=3)
```

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.854</td>
<td>0.098</td>
<td>-8.697</td>
<td>0.000</td>
</tr>
<tr>
<td>pulse.cent</td>
<td>0.016</td>
<td>0.008</td>
<td>1.948</td>
<td>0.051</td>
</tr>
</tbody>
</table>
### NHANES: Pulse Rate & Sleep

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.854</td>
<td>0.098</td>
<td>-8.697</td>
<td>0.000</td>
</tr>
<tr>
<td>pulse.cent</td>
<td>0.016</td>
<td>0.008</td>
<td>1.948</td>
<td>0.051</td>
</tr>
</tbody>
</table>

- **Slope**: As pulse rate increases by 1 beat per minute, the odds of having sleep trouble are expected to multiply by a factor of 1.02, with 95% confidence interval 1 to 1.032
NHANES: Pulse Rate & Sleep

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.854</td>
<td>0.098</td>
<td>-8.697</td>
<td>0.000</td>
</tr>
<tr>
<td>pulse.cent</td>
<td>0.016</td>
<td>0.008</td>
<td>1.948</td>
<td>0.051</td>
</tr>
</tbody>
</table>

- **Slope:** As pulse rate increases by 1 beat per minute, the odds of having sleep trouble are expected to multiply by a factor of 1.02, with 95% confidence interval 1 to 1.032

- **Intercept:** A person with an average pulse (72.452 beats per minute) is expected to have 0.43 to 1 odds of having sleep trouble, with 95% confidence interval 0.35 to 0.515
NHANES: Computers & Sleep

- **CompHrsDay**: Number of hours per day on average participant used a computer or gaming device over the last 30 days.
  - 0_hrs
  - 0_to_1_hr
  - 1_hr
  - 2_hr
  - 3_hr
  - 4_hr
  - More_4_hr

- **Age**: Age at time of screening (in years). Participants 80 or older were recorded as 80.
### NHANES: Computers & Sleep

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.168</td>
<td>0.594</td>
<td>-3.648</td>
<td>0.000</td>
</tr>
<tr>
<td>CompHrsDay0_to_1_hr</td>
<td>0.284</td>
<td>0.382</td>
<td>0.745</td>
<td>0.456</td>
</tr>
<tr>
<td>CompHrsDay1_hr</td>
<td>0.436</td>
<td>0.419</td>
<td>1.040</td>
<td>0.298</td>
</tr>
<tr>
<td>CompHrsDay2_hr</td>
<td>0.412</td>
<td>0.530</td>
<td>0.776</td>
<td>0.438</td>
</tr>
<tr>
<td>CompHrsDay3_hr</td>
<td>1.331</td>
<td>0.650</td>
<td>2.047</td>
<td>0.041</td>
</tr>
<tr>
<td>CompHrsDay4_hr</td>
<td>1.803</td>
<td>0.990</td>
<td>1.822</td>
<td>0.068</td>
</tr>
<tr>
<td>CompHrsDayMore_4_hr</td>
<td>0.318</td>
<td>0.626</td>
<td>0.508</td>
<td>0.611</td>
</tr>
<tr>
<td>Age</td>
<td>0.024</td>
<td>0.009</td>
<td>2.590</td>
<td>0.010</td>
</tr>
</tbody>
</table>

- Describe the relationship between the amount of time spent on a computer and the odds of having trouble sleeping after adjusting for age.
Assumptions & Model Fit

- **Assumptions**
  - Independence of residuals
  - Log-odds has linear relationship with explanatory variables
  - No significant impacts due to influential points or multicollinearity (when there are multiple explanatory variables)
Assumptions & Model Fit

- **Assumptions**
  - Independence of residuals
  - Log-odds has linear relationship with explanatory variables
  - No significant impacts due to influential points or multicollinearity (when there are multiple explanatory variables)

- Not effective to examine the residuals
  - Residual positive when $Y = 1$ and negative when $Y = 0$
  - Constant variance not an assumption of logistic regression
  - Normality of residuals not an assumption of logistic regression
Assumptions & Model Fit

- **Check Assumptions**
  - Plot of binned residuals vs. predicted values
  - Plot of binned residuals vs. numeric explanatory variables
  - Leverage, Cook's distance, and multicollinearity (if more than one explanatory variable)
Assumptions & Model Fit

- **Check Assumptions**
  - Plot of binned residuals vs. predicted values
  - Plot of binned residuals vs. numeric explanatory variables
  - Leverage, Cook's distance, and multicollinearity (if more than one explanatory variable)

- **Check Model Fit**
  - Examine ROC curve
  - Examine confusion matrix
Binned Residuals

To examine binned residuals...

- Calculate raw residuals

- Order observations either by the values of the predicted probabilities or by the numeric explanatory variable itself

- Use the ordered data to create $g$ bins of approximately equal size
  - Default value: $g = \sqrt{n}$

- Calculate average residual value in each bin

- Plot average residuals vs. average predicted probability (or average explanatory variable)
Binned Residuals

To examine binned residuals...

- Calculate raw residuals
- Order observations either by the values of the predicted probabilities or by the numeric explanatory variable itself
- Use the ordered data to create \( g \) bins of approximately equal size
  - Default value: \( g = \sqrt{n} \)
- Calculate average residual value in each bin
- Plot average residuals vs. average predicted probability (or average explanatory variable)

Can use the \texttt{arm} package in R
Binned Residuals

- Look for patterns

- Nonlinear trend may be indication that squared term or log transformation of explanatory variable required

- If bins have average residuals with large magnitude
  - Look at averages of other explanatory variables across bins
  - Interaction may be required if large residuals correspond to certain combinations of explanatory variables
Raw Residuals: Pulse & Sleep

```r
nhanes <- nhanes %>%
  mutate(Residuals = residuals.glm(log.odds,type="response"),
         Predicted = predict.glm(log.odds,type="response"))
```
Binned Residuals - Pulse & Sleep

```r
library(arm)
binnedplot(x=nhanes$Predicted,y=nhanes$Residuals,xlab="Predicted Pr
```
Binned Residuals - Pulse & Sleep

```r
library(arm)
binnedplot(x=nhanes$Pulse, y=nhanes$Residuals, xlab="Pulse")
```