Lab 11: Multiple Logistic Regression

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> ”The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.”
>
> *Stephen Hawking*

Introduction

The basic idea of Lab is as follows: Answer a research question with the provided dataset. Each week, that research question (and data) will change depending on the topic we've covered the prior class days. Once we're done with Lab, you'll have a Lab Assignment, that will look a lot like the Lab: a research question you'll need to answer given some data. In Lab, you'll learn the procedure for answering the research question. For the Lab Assignment, you'll do that procedure for a grade (independently).

To help answer the research question, we'll follow some basic steps that we'll repeat throughout the semester:

- Reflect on the Question: Figure out the variables of interest, and the technique that's required.
- Analyze the Data: Perform the steps required for the technique.
- Draw Conclusions: Use the information that you got from the prior step to answer the research question in a concise, logical manner.

Let's get started:

Primary Research Question:

What is the best predictor of Low Birth Weight among mothers: age, mother's weight, or smoking status? Explain and graph your findings.

Step1: Reflect on the Question:

Download the syntax and data files from Canvas.

Let's load in our SDSRegressionR package so that we can use some of it's functions later:

Remember we need the *newest* version of the package for this lab.

```
#Load our class package
library(SDSRegressionR)
```

Next, we'll load in the data. Be sure to use the basic file structure we talked about the first Lab: Put your syntax in a folder specific to this Lab. Then, make a “data” folder in that same place - use lowercase. If you do that, then all of this syntax will work like a charm.

```r
birth <- read_csv("data/LowBirth.csv")
```

Check the Data:

To make sure that we're working with the right data, and that we're all looking at it the same way, we'll answer some basic questions about the data before moving on:

1. How many observations are in the dataset for the model?

2. What was the Age for the mother with the first Low Birth Weight observation?

3. Of the first 10 observations, mothers were smokers?

These questions can be answered simply by looking at the dataset once it's loaded in:

```r
View(birth)
```

Check the Variables of Interest

Let's find the variables that we need to answer the primary research question:

1. Which variable tells us the Outcome variable for this model?
   - What type of variable is this?

2. What are the variables of interest for the model?

3. Classify each of the variables of interest.

Again, these can be answered by looking at the dataframe, and with the help of the *names()* function. Also, the codebook for the data frame is our friend. You can open this in R or Excel. Remember, R is case-sensitive.

```r
names(birth)
```

```
##  [1] "ID"    "LOW"   "AGE"   "LWT"   "RACE"  "SMOKE" "PTL"   "HT"
##  [9] "UI"    "FTV"   "BWT"
```

Reflect on the Method

The last part of Reflect on the Question asks about the method or technique we'll use.

1. We will use Multiple Logistic Regression to answer this question. Why?

2. We'll need to run vif() for this model. Why?

3. We'll use a graph of predicted probabilities. Why?

Step2: Analyze the Data

In this step, we'll run the provided syntax and answer some questions about the output to help us prepare for the final step.

Here's the syntax you'll need (from the .R syntax file):

```r
#### Here is the R script you will use:  (remember that # indicates a comment) ####
#Lab11: Multiple Logistic Regression

library(SDSRegressionR)
#Bring in data
birth <- read.csv("data/LowBirth.csv", stringsAsFactors = FALSE)
names(birth)

#Intially:
table(birth$LOW)

#Factoring the categorical variables in the model (but NOT the outcome):
table(birth$SMOKE)
birth <- birth %>%
  mutate(SMOKE_f = factor(SMOKE, levels=c(0,1),
                         labels=c("Non-smoker", "Smoker")))

#Intital Model
b_mod <- glm(LOW ~ AGE + LWT + SMOKE_f, data=birth, family="binomial")
summary(b_mod)

library(car)
vif(b_mod)
cooksPlot(b_mod, key.variable = "ID", print.obs = TRUE, sort.obs = TRUE)
threeOuts(b_mod, key.variable = "ID")

#Get good data...
g_birth <- birth %>%
  filter(ID %not in% c(21, 140, 11, 159))

#Re-run
b_mod2 <- glm(LOW ~ AGE + LWT + SMOKE_f, data=g_birth, family="binomial")
summary(b_mod2)

#Odds-ratios
exp(b_mod2$coef)
exp(confint.default(b_mod2))

#Stats
library(rms)
b_mod2.2 <- lrm(LOW ~ AGE + LWT + SMOKE_f, g_birth)
b_mod2.2

#Examine the variables of interest graphically...
#Look at ranges...
library(skimr)
g_birth %>%
  skim(LWT)

#Predict
library(emmeans)
birth_mns <- summary(emmeans(b_mod2, "LWT",
                             at=list(LWT = seq(80, 250, 10)), type="response"))
birth_mns

#Graph
g <- simpleScatter(g_birth, LWT, LOW, title="Low birth weight",
                   xlab="Mother's weight", ylab="Low birth weight probability")
g
g +
  geom_line(data=birth_mns, aes(x=LWT, y=prob), color="red") +
  geom_line(data=birth_mns, aes(x=LWT, y=asymp.LCL), linetype="dashed") +
  geom_line(data=birth_mns, aes(x=LWT, y=asymp.UCL), linetype="dashed")

#Out of curiosity...
birth_mns_smk <- summary(emmeans(b_mod2, "SMOKE_f", by="SMOKE_f", type="response"))
birth_mns_smk

#Relative Risk
0.3391334 / 0.2318478

ggplot(birth_mns_smk, aes(y=prob, x=factor(SMOKE_f))) +
  geom_point(size=2) +
  geom_errorbar(aes(ymin=asymp.LCL, ymax=asymp.UCL), width=.1) +
  ylim(0,1) +
  labs(title="Smoking and low birth", subtitle="With 95% CI",
       x="Smoker Status", y="Probability of Low Birth Weight") +
  geom_hline(yintercept = 0.5, color="red") +
  theme_bw()

##EXTRA
birth_mns2 <- summary(emmeans(b_mod2, "LWT",
                              at=list(LWT = seq(80, 250, 1)), type="response"))
mark_prob <- 0.25
ci_marks <- birth_mns2 %>%
  filter(abs(asymp.LCL - mark_prob) == min(abs(asymp.LCL - mark_prob)) |
           abs(asymp.UCL - mark_prob) == min(abs(asymp.UCL - mark_prob))) %>%
  pull(LWT)

g +
  geom_line(data=birth_mns, aes(x=LWT, y=prob), color="red") +
  geom_line(data=birth_mns, aes(x=LWT, y=asymp.LCL), linetype="dashed") +
  geom_line(data=birth_mns, aes(x=LWT, y=asymp.UCL), linetype="dashed") +
  geom_vline(xintercept = ci_marks)
```

All quesitons are in reference to models run *after* outlier removal.

Question 1

The overall model was significant in the prediction of Low Birth Weight: LR chi2 (_______) = _______, p < 0.05.

Question 2

Pseudo R^2 showed _______ of the variance in Low Birth Weight could be accounted for by the predictors.

Question 3

Of the three predictors, _______ was significantly related to Low Birth Weight status (OR = ______). Report the meaningful statistics: z = _______, p = _______.

Question 4

Neither mother's Age nor Smoking Status were significant predictors of Low Birth Weight (Age: OR= _______, z = _______, p = _______; Smoking: OR= _______, z = _______, p = _______).

Bonus:

At what Mother's Weight will there be a 0.25 probability of a Low Birth classification?

Step3: Draw Conclusions

The final step is for us to Draw Conclusions. We'll take the syntax we've been given from Analyze the Question, run it, then examine the output. The questions from the prior step help set us up for the Draw Conclusions part.

We'll "fill in the blanks" in a canned paragraph for the Lab. For the Lab Assignment, you'll need to come up with a similar paragraph all on your own (please don't steal mine).

Primary Research Question
What is the best predictor of Low Birth Weight among mothers: age, mother's weight, or smoking status? Explain and graph your findings.

> Multiple Logistic Regression analysis was used to examine the predictors of Mother Age, Mother's Weight at last menstruation, and Mother's Smoking Status on Low Birth Weight classification. The model was significant overall (LR chi2(3) = _______, $p < 0.05$), accounting for _______ of the variance in the outcome (Pseudo R2). Of the three predictors in the model, only Weight of the mother was significant (OR = _______, z = _______, $p < 0.05$). However, although not significant, Smoking status had a large Odds Ratio (OR = _______, z = _______, p = _______), indicating that a transition to smoking status increased the odds of a Low Birth Weight classification 1.70 times. However, Releative Risk calculation shows this effect to be smaller at _______. A non-smoking mother has a predicted probability of _______ (SE: _______), while a smoking mother has a predicted probability of _______, (SE: _______).

Lab Assignment

Now, with the tools at your disposal (the R syntax from Lab, and the logic of proceeding through the three steps of answering the research question), you'll have a Lab Assignment to complete (independently). For now, the Lab Assignment is to be completed in Canvas. It will follow the basic structure, and lead to the same place - answering the research question with a concise paragraph as in Draw Conclusions.

Good Luck!