# Logistic Regression Modeling Using APGAR Score and Toxemia as Predictor on Germinal Matrix Hemorrhage

K Boswell

2023-08-04

## Background

The dataset lowbwt.sas7bdat contains information for the sample of 100 low birth weight infants born in Boston, Massachusetts. The variable grmhem is a dichotomous random variable indicating whether an infant experienced a germinal matrix hemorrhage. The value 1 indicates that a hemorrhage occurred and 0 that it did not. The infants' five-minute apgar scores are saved under the name apgar5, and indicators of toxemia – where 1 represents a diagnosis of toxemia during pregnancy for the child's mother and 0 no such diagnosis – under the variable name tox. This markdown file contains the approach I took to solving a series of questions related to the model.

### *Step 1: Read the data and view the structure:*

```
library(haven)
lowbwt <- read_sas("lowbwt.sas7bdat")

str(lowbwt)
```

```
## tibble [100 × 6] (S3: tbl_df/tbl/data.frame)
##  $ sbp    : num [1:100] 43 51 42 39 48 31 31 40 57 64 ...
##   ..- attr(*, "label")= chr "sbp"
##  $ sex    : num [1:100] 1 1 0 0 0 1 1 0 0 0 ...
##   ..- attr(*, "label")= chr "sex"
##  $ tox    : num [1:100] 0 0 0 0 1 0 1 0 0 1 ...
##   ..- attr(*, "label")= chr "tox"
##  $ grmhem : num [1:100] 0 0 0 0 0 1 0 0 0 0 ...
##   ..- attr(*, "label")= chr "grmhem"
##  $ gestage: num [1:100] 29 31 33 31 30 25 27 29 28 29 ...
##   ..- attr(*, "label")= chr "gestage"
##  $ apgar5 : num [1:100] 7 8 0 8 7 0 7 9 6 9 ...
##   ..- attr(*, "label")= chr "apgar5"
```

The structure of the data indicates some problems. There are no na values, which is good, but the categorical variables are all stored as numbers instead of factors.

### *Step 2: Clean and review the data*

Note: The data currently exists as a tibble instead of a data frame. I will convert the tibble into a data frame to correspond to my current knowledge of R structures. I also will convert the numbered data into factors. Then I will double-check that no columns are missing data and everything has been converted as I expected.

```
# The structure shows the data is in a tibble. Find the tibble name:
tibble_name <- deparse(substitute(lowbwt))
print(tibble_name)
```

```
## [1] "lowbwt"
```

```
# Convert tibble into dataframe:
lowbwt_df <- as.data.frame(lowbwt)

# Convert numbered variables into binary for tox and grmhem
# Convert tox, sex, and grmhem into F and M for easier reading and convert column into a factor:
lowbwt$tox <- as.factor(lowbwt$tox)
lowbwt$sex <- as.factor(lowbwt$sex)
lowbwt$grmhem <- as.factor(lowbwt$grmhem)

# Check the data structure
str(lowbwt)
```

```
## tibble [100 × 6] (S3: tbl_df/tbl/data.frame)
##  $ sbp    : num [1:100] 43 51 42 39 48 31 31 40 57 64 ...
##   ..- attr(*, "label")= chr "sbp"
##  $ sex    : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 1 1 ...
##  $ tox    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 1 2 ...
##  $ grmhem : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ gestage: num [1:100] 29 31 33 31 30 25 27 29 28 29 ...
##   ..- attr(*, "label")= chr "gestage"
##  $ apgar5 : num [1:100] 7 8 0 8 7 0 7 9 6 9 ...
##   ..- attr(*, "label")= chr "apgar5"
```

```
# Check if there are any missing values in each column
col_has_missing <- colSums(is.na(lowbwt)) > 0
print(col_has_missing)
```

```
##     sbp     sex     tox  grmhem gestage  apgar5 
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
```

```
nrow(lowbwt)
```

```
## [1] 100
```

## *Step 3: Create Tables to view the data as two-way contingency tables*

```
xtabs(~ grmhem + tox, data=lowbwt)
```

```
##       tox
## grmhem  0  1
##      0 65 20
##      1 14  1
```

```
xtabs(~ grmhem + apgar5, data = lowbwt)
```

```
##       apgar5
## grmhem  0  1  2  3  4  5  6  7  8  9
##      0  5  1  1  3  3  7  9 21 22 13
##      1  1  0  2  1  2  3  2  2  2  0
```

There is some imbalanced data with only 1 patient experiencing germinal matrix hemorrhage with toxemia. This is a potential problem to the model, but it is not something to deal with right now. For now, I will go forward with modeling and just keep this in mind.

## *Step 4: Begin Modeling*

Question 3a: Determine an equation for a logistic regression model where germinal matrix hemorrhage is the response and five-minute APGAR is the predictor, using β1 to represent the regression coefficient of apgar score.

Answer: log{Pr(grmhem=1|apgar5) / (1 - Pr(grmhem=1|apgar5))} = β0 + β1 * apgar5

# Question 3b: Fit the logistic regression model in part (a). What is βˆ1, the estimated regression coefficient of apgar score? What's the interpretation of βˆ1?

```
# Fit the logistic regression model
model <- glm(grmhem ~ apgar5, data = lowbwt, family = "binomial")

# View the summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = grmhem ~ apgar5, family = "binomial", data = lowbwt)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3037     0.6191  -0.491   0.6237
## apgar5       -0.2496     0.1044  -2.392   0.0168 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 84.542  on 99  degrees of freedom
## Residual deviance: 78.927  on 98  degrees of freedom
## AIC: 82.927
##
## Number of Fisher Scoring iterations: 4
```

Answer: The coefficient of apgar5 is -.2496. This represents the change in the log odds of experiencing germinal matrix hemorrhage for each one-unit increase in the five-minute apgar score. To interpret this, we need to exponentiate -.2496, which is .7791. That means that the log odds of experiencing the grmhem decrease by .7791 for every one-unit increase in the apgar score.

Question 3c: If a particular child has a five-minute apgar score of 3, what is the predicted probability that this child will experience a brain hemorrhage?

Answer: We need to calculate the log odds for an apgar5 score of 3 and then calculate the probability of experiencing a hemorrhage for an apgar5 score of 3.The formula for the probability is p <- exp(log_odds) / (1 + exp(log_odds)).

```
# Extract the coefficient for apgar 5
coef_apgar5 <- coef(model)["apgar5"]
coef_beta0 <- coef(model)["(Intercept)"]

# Calculate the probability of grmhem for a specific value of apgar5
apgar5_score_3 <- 3
log_odds_score3 <- exp(coef_beta0 + 3*coef_apgar5)
probability_3 <- exp(coef_beta0 + 3*coef_apgar5)/(1 + exp(coef_beta0 + 3*coef_apgar5))

# Print the results
print(paste("Probability of Germinal Matrix Hemorrhage for APGAR5 score of", apgar5_score_3, " =
", probability_3))
```

```
## [1] "Probability of Germinal Matrix Hemorrhage for APGAR5 score of 3  =  0.258735085840488"
```

The predicted probability is 25.87% that if a particular child has a five-minute apgar score of 3, that he or she will experience a brain hemorrhage.

Question 3d: What is the estimated odds ratio of suffering a germinal matrix hemorrhage associated with 1 unit increase in five-minute apgar score?

Answer: Using the coefficient of apgar5 = -.2496, we get the odds ratio by exponentiating:

```
# Get the odds ratio
odds_ratio_apgar5 <- exp(coef_apgar5)

# Print the results
print(paste("Odds Ratio for APGAR5:", odds_ratio_apgar5))
```

```
## [1] "Odds Ratio for APGAR5: 0.779106571657109"
```

The odds ratio is approximately .7791, meaning that for each one-unit increase in the apgar5 score, the odds of experiencing a germinal matrix hemorrhage decrease by approximately 22.1% (100 - 77.9 = 22.1). Or put another way, for each one-unit decrease in the apgar5 score, the odds of experiencing a germinal matrix hemorrhage increase by approximately 22.1%.

Question 3e: What is the estimated odds ratio of suffering a germinal matrix hemorrhage associated with 3 units increase in five-minute apgar score?

```
# Calculate the log odds for a value of apgar5_score_3 + 3 (3 units increase)
apgar5_score_3_plus_3 <- apgar5_score_3 + 3
log_odds_score3_plus_3 <- coef_apgar5 * apgar5_score_3_plus_3

# Calculate the odds ratio for a 3-unit increase in apgar5
odds_ratio_3_units_increase <- exp(log_odds_score3_plus_3 - log_odds_score3)

# Print the results
print(paste("Odds Ratio for a 3-unit increase in APGAR5 score =", odds_ratio_3_units_increase))
```

```
## [1] "Odds Ratio for a 3-unit increase in APGAR5 score = 0.157758485223081"
```

Answer: to calculate the odds ratio associated with a 3-unit increase, we take the odds ratio value of .7791 and raise it to the power of 3. .7791^3 = .473. We can also calculate it as I did above. The results show that for every 3-unit incrase in the apgar5 score, the odds of experiencing a germinal matrix hemorrhage decrease by approximately 52.7% (100% - 47.3%). Or put another way, for each 3-unit decrease in the apgar 5 score, the odds of experiencing a germinal matrix hemorrhage increase by approximatly 52.7%.

Question 3f: Write down the equation for a logistic regression model where germinal matrix hemorrhage is the response and toxemia status is the predictor, using β2 to represent the regression coefficient of toxemia status.

Answer: log{Pr(grmhem=1|tox) / (1 - Pr(grmhem=1|tox))} = β0 + β2 * tox

Question 3g: Fit the logistic regression model. What is βˆ2, the estimated regression coefficient of toxemia status? What’s the interpretation of βˆ2?

```
# Fit the logistic regression model
tox_model <- glm(grmhem ~ tox, data = lowbwt, family = binomial)

# View the summary of the model
summary(tox_model)
```

```
##
## Call:
## glm(formula = grmhem ~ tox, family = binomial, data = lowbwt)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5353     0.2946  -5.211 1.88e-07 ***
## tox1         -1.4604     1.0661  -1.370    0.171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 84.542  on 99  degrees of freedom
## Residual deviance: 81.849  on 98  degrees of freedom
## AIC: 85.849
##
## Number of Fisher Scoring iterations: 5
```

Answer: The coefficient of tox is -.1.4604. This represeents the change in the log odds of experiencing germinal matrix hemorrhage for individuals with toxemia compared to those without toxemia. Since the coefficient is negative, it means the the log odds of experiencing a germinal matrix hemorrhage decrease by 1.4604 for individuals with toxemia compared to those without.However, the coefficient is not statistically significant, indicating that there is no confidence in this interpretation. In addition, this is where I went back to the contingency table to reflect on the outcomes. Given that only 1 patient experienced germinal matrix hemorrhage who had toxemia, I would question my sample and go back to the data to see if I could pull a more balanced sample. If the data does not provide a more balanced sample, we are stopped here. For the data provided in this exercise, I would fail to reject the null hypothesis.

Question 3h: For a child whose mother was diagnosed with toxemia during pregnancy, what is the predicted probability of experiencing a germinal matrix hemorrhage?

Answer: First, I don’t think this dataset actually is good for prediction given the non-significance of tox1. However, if we still wanted to use the data to calculate the log odds for having toxemia and then calculate the probability of experiencing a hemorrhage for given that the patient has toxemia, then the formula for the probability is p <- exp(log_odds) / (1 + exp(log_odds)).

```
# Get the coefficient for tox1 (B1)
coef_beta0_tox <- coef(tox_model)["(Intercept)"]
coef_tox1 <- coef(tox_model)["tox1"]

# Calculate the probability (p) of experiencing a germinal matrix hemorrhage for tox = 1
log_odds_tox <- exp(coef_beta0_tox + coef_tox1)
probability_tox1 <- exp(coef_beta0_tox + coef_tox1) / (1 + exp(coef_beta0_tox + coef_tox1))

# Display the predicted probability
# Print the results
print(paste("Probability of Germinal Matrix Hemorrhage for Toxemia = True is", " = ", probabilit
y_tox1))
```

```
## [1] "Probability of Germinal Matrix Hemorrhage for Toxemia = True is  =  0.04761904819819"
```

Therefore, the predicted probability of a patient with toxemia experiencing a germinal matrix hemorrhage is 4.76%. Again, however, since the coefficient for tox1 was not significant, the reliability of this prediction is low.

Question 3i: What are the estimated odds of suffering a germinal matrix hemorrhage for children whose mothers were diagnosed with toxemia relative to children whose mothers were not?

Answer: we need the odds ratio of the toxemia coefficient.

```
# Get the coefficient for tox1
coeff_tox1 <- coef(tox_model)["tox1"]

# Calculate the estimated odds ratio (OR)
estimated_OR <- exp(coeff_tox1)

# Display the estimated odds ratio
estimated_OR
```

```
##      tox1
## 0.2321429
```

Since this odds ratio is lower than 1, it means that the odds are about .2321 lower for mothers with toxemia than for mothers without toxemia. Again, though, since the initial data is imbalanced, and the p-value of β2 is low, I do not have confidence in this conclusion.