**Project:** Project 2, BIOS 6623

**Report:** VA Surgery Database Review

**Investigator:** Rachel Johnson

**Date:** 1 November 2017

**Introduction**

The goal of this project is to evaluate 30 day mortality rates among 26,518 patients who have undergone heart surgery at 44 VA hospitals in the most recent six-month period, period 39. Certain demographics were collected among these patients, including hospital code; the six month period in which they had their surgery; procedure type—CABG or valve surgery; height; weight; BMI; ASA—a code for the patient’s condition at the start of surgery; albumin levels; and 30 day mortality. Observed and expected mortality rates were used to determine which hospitals’ mortality rates were clinically and/or statistically unusual and warranted a site visit. The purpose of these site visits would be to either identify potential problems if the rates were unusually high or to perform a site visit to determine if there is anything to learn from in order to improve outcomes at other hospitals if the rates were unusually low.

**Methods**

I was assigned to analyze VA Data 2.

There were originally 26,420 individuals from 44 hospitals in the data set. 2 individuals were removed because they underwent surgeries other than valve surgery or CABG surgery, so they should not have been included in the data set.

There were issues with weight entry in hospitals 1-16 in the most recent period, as weights were entered in kilograms. Other BMIs in the data set were plotted against the calculated BMIs, and any incorrectly entered or calculated BMIs were corrected individually.

The variable ASA, which is an indicator of the patient’s condition at the start of surgery was dichotomized such that ratings of 1, 2, and 3 which indicate better health were in one category, while ratings of 4 or 5 were in another category, which indicated worse health at the beginning of the surgery.

Demographic information at the overall level was summarized in Table 1, with continuous variables summarized with mean and standard deviation and categorical variables summarized with frequency and percentages. Frequencies of missing data were also noted for each variable, if applicable.

13,239 individuals (49.9%) of the total population were missing values for albumin. There were no obvious differences in other data set values between those with albumin values and for those without albumin values as determined through marginal plots and frequency tables that would have explained the missingness.

30 day mortality for individuals was modeled in a logistic regression that accounted for procedure type, ASA, and BMI. This regression was also modeled with albumin as a predictor to compare estimates and to determine if albumin would be a valuable measure to collect for all patients in the future.

The estimates from this model were used to determine fitted values for each individual in the last six month period for which all covariate data was available, which were then averaged by hospital to determine expected mortality rates for each hospital in period 39. Hospital 30 had no individuals with all covariate data available since they were all missing BMI values in period 39, so no expected mortality rates could be calculated for this hospital.

These expected mortality rates were compared to the observed mortality rates for each hospital in this period—excluding hospital 30. If the ratio of observed to expected mortality rates was greater than 1.2, they were deemed unusually high by the clinical definition, which would warrant a site visit in order to identify and resolve any potential problems. If the ratio of observed to expected mortality rates was less than 0.8, a site visit would also be planned to determine if there is anything to learn from in order to improve outcomes at other hospitals if the rates were unusually low.

To determine an interval of variation around the average expected probability of mortality by hospital, bootstrap confidence intervals were required. Data was resampled from the complete cases data set, which included individuals who had values for hospital code, six month period, BMI, ASA, procedure type, and mortality. A logistic regression was run on this resampled data, and the estimates were used to find predicted probabilities of mortality on the original data set from period 39, and then were averaged by hospital. There were 10,000 iterations run in this bootstrap, and then confidence intervals for the average expected probability of mortality were determined by hospital by calculating the 2.5th and 97.5th percentile of the distribution for each hospital’s average.

**Results**

The demographics of this data set are shown in Table 1, and the only variable with concerning levels of missing data is albumin, with 49.9% missingness. The concern is due to the bias that arises since those who are more likely to be missing covariate data had much higher rates of death than those who were not missing any covariate data. Those were not missing any covariate data had a death rate of 2.89% in period 39, while those who were missing covariate data had a death rate of 8.41%. Of these with missing covariate data, there was an 8.01% mortality rate for those with missing BMI in period 39, and there was a 9.71% mortality rate for those missing procedure data in period 39. Therefore, since this missingness is related to the outcome, this is likely causing bias that should be considered when analyzing the remainder of the results since expected rates will likely be too low.

**Table 1.** This table displays the summary of the VA data on heart surgeries at 44 hospitals over the last 6 periods.

|  |  |
| --- | --- |
| **Hospitals (n)** | 44 |
| **Patients undergoing heart surgery (n)** | 26518 |
| **Procedure (n (%))** |  |
| Valve surgery | 5610 (21.16) |
| CABG surgery | 21457 (80.91) |
| **ASA (n (%))** |  |
| 3 or less | 6296 (23.74) |
| 4 or greater | 19558 (73.75) |
| Missing | 664 (2.5) |
| **BMI (mean (SD))** | 28.64 (3.78) |
| Missing (n (%)) | 702 (2.65) |
| **Albumin (mean (SD))** | 4.02 (0.55) |
| Missing (n (%)) | 13239 (49.92) |
| **30 day mortality (n (%))** | 834 (3.15) |

The results of the logistic regression modeling individuals’ 30 day mortality are shown in Table 2, and this results of this logistic regression that was run without albumin are in Table 3. Procedure and ASA were significant predictors of 30 day mortality in both models, and BMI was not significant in either model. The addition of albumin was not a significant predictor in this model, and did not dramatically change the estimates or directions of the other three predictors. Therefore, it was not included in this model so as not to dramatically reduce the sample size, and it likely does not need to be collected on all patients in the future so that it could be added to models, since it does not seem to be a valuable predictor.

**Table 2.** This table displays the results of the logistic regression with 30 day mortality as the outcome, with odds ratios, 95% confidence intervals, and p-values for each of the four predictors.

|  |  |  |  |
| --- | --- | --- | --- |
| **Covariate** | **Estimate** | **95% CI** | **p-value** |
| Procedure (reference = valve surgery) | 1.337 | 1.002, 1.785 | 0.048 |
| ASA (reference = 3 or less) | 2.379 | 1.681, 3.367 | <0.001 |
| BMI | 0.997 | 0.968, 1.026 | 0.837 |
| Albumin | 0.952 | 0.772, 1.175 | 0.649 |

**Table 3.** This table displays the results of the logistic regression with 30 day mortality as the outcome, with odds ratios, 95% confidence intervals, and p-values for each of the three predictors, once albumin was removed.

|  |  |  |  |
| --- | --- | --- | --- |
| **Covariate** | **Estimate** | **95% CI** | **p-value** |
| Procedure (reference = valve surgery) | 1.384 | 1.128, 1.698 | 0.002 |
| ASA (reference = 3 or less) | 2.707 | 2.127, 3.447 | <0.001 |
| BMI | 1.009 | 0.988, 1.029 | 0.407 |

The observed and expected proportions of death for each hospital are shown in Table 4, along with the 95% confidence intervals obtained from the bootstrap. It is noted which hospitals’ observed rates were higher than the upper bound of the 95% confidence intervals, of which there are 17.

Furthermore, it was clinically determined that if the ratio of the observed percent died to the predicted percent died for the last 6 months was greater than 1.2 or less than 0.8 that they were unusually high or unusually low, and warranted a site visit from the VA. There were 17 hospitals whose observed percent to expected percent died were unusually high, and there were 22 hospitals whose observed percent to expected percent died were unusually low, based on these definitions.

**Table 4.** This table displays the mortality rates at each hospital, along with the predicted mortality rates based on the logistic regression fitted values, the bootstrap 95% confidence intervals, and whether these observed percentages were unusually high or low in reference to the predicted and bootstrapped values.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hospital** | **Patients died** | **Patients in surgery** | **Percent died** | **Predicted percent died for last 6 months (population-adjusted)** | **Bootstrap 95% CI** | **Observed Above Confidence Interval?** | **Unusually High?** | **Unusually Low?** |
| 1 | 1 | 87 | 1.15 | 3.02 | 2.88, 3.31 | No | No | Yes |
| 2 | 1 | 106 | 0.94 | 3.17 | 3.01, 3.47 | No | No | Yes |
| 3 | 4 | 100 | 4 | 3.1 | 2.97, 3.41 | Yes | Yes | No |
| 4 | 4 | 94 | 4.26 | 3.02 | 2.91, 3.34 | Yes | Yes | No |
| 5 | 1 | 115 | 0.87 | 3.14 | 2.98, 3.43 | No | No | Yes |
| 6 | 2 | 104 | 1.92 | 2.92 | 2.83, 3.25 | No | No | Yes |
| 7 | 7 | 105 | 6.67 | 2.84 | 2.75, 3.17 | Yes | Yes | No |
| 8 | 4 | 120 | 3.33 | 3.13 | 2.99, 3.45 | No | No | No |
| 9 | 0 | 105 | 0 | 2.93 | 2.81, 3.24 | No | No | Yes |
| 10 | 2 | 100 | 2 | 2.94 | 2.83, 3.25 | No | No | Yes |
| 11 | 1 | 90 | 1.11 | 2.95 | 2.82, 3.25 | No | No | Yes |
| 12 | 4 | 98 | 4.08 | 3.09 | 2.95, 3.39 | Yes | Yes | No |
| 13 | 4 | 84 | 4.76 | 3.07 | 2.93, 3.36 | Yes | Yes | No |
| 14 | 1 | 103 | 0.97 | 2.86 | 2.78, 3.2 | No | No | Yes |
| 15 | 3 | 105 | 2.86 | 3.03 | 2.9, 3.34 | No | No | No |
| 16 | 1 | 111 | 0.9 | 2.96 | 2.84, 3.27 | No | No | Yes |
| 17 | 13 | 93 | 13.98 | 3.15 | 3, 3.45 | Yes | Yes | No |
| 18 | 2 | 95 | 2.11 | 3.02 | 2.89, 3.33 | No | No | Yes |
| 19 | 0 | 113 | 0 | 2.96 | 2.85, 3.27 | No | No | Yes |
| 20 | 2 | 98 | 2.04 | 2.75 | 2.67, 3.09 | No | No | Yes |
| 21 | 5 | 92 | 5.43 | 3.04 | 2.91, 3.35 | Yes | Yes | No |
| 22 | 2 | 86 | 2.33 | 3.07 | 2.93, 3.37 | No | No | Yes |
| 23 | 6 | 97 | 6.19 | 2.83 | 2.74, 3.15 | Yes | Yes | No |
| 24 | 4 | 104 | 3.85 | 2.98 | 2.86, 3.29 | Yes | Yes | No |
| 25 | 3 | 95 | 3.16 | 2.99 | 2.87, 3.3 | No | No | No |
| 26 | 4 | 99 | 4.04 | 2.94 | 2.82, 3.25 | Yes | Yes | No |
| 27 | 2 | 99 | 2.02 | 3.09 | 2.93, 3.38 | No | No | Yes |
| 28 | 5 | 101 | 4.95 | 3.2 | 3.04, 3.49 | Yes | Yes | No |
| 29 | 2 | 105 | 1.9 | 2.94 | 2.83, 3.25 | No | No | Yes |
| 30 | 10 | 117 | 8.55 | NA | NA | NA | NA | NA |
| 31 | 7 | 104 | 6.73 | 2.93 | 2.82, 3.25 | Yes | Yes | No |
| 32 | 0 | 93 | 0 | 2.99 | 2.88, 3.31 | No | No | Yes |
| 33 | 0 | 113 | 0 | 3.02 | 2.9, 3.33 | No | No | Yes |
| 34 | 14 | 99 | 14.14 | 3.01 | 2.89, 3.32 | Yes | Yes | No |
| 35 | 5 | 84 | 5.95 | 2.92 | 2.81, 3.23 | Yes | Yes | No |
| 36 | 1 | 99 | 1.01 | 2.84 | 2.75, 3.18 | No | No | Yes |
| 37 | 4 | 107 | 3.74 | 2.89 | 2.79, 3.21 | Yes | Yes | No |
| 38 | 1 | 113 | 0.88 | 2.9 | 2.79, 3.21 | No | No | Yes |
| 39 | 4 | 101 | 3.96 | 3 | 2.88, 3.32 | Yes | Yes | No |
| 40 | 2 | 86 | 2.33 | 2.94 | 2.82, 3.24 | No | No | Yes |
| 41 | 5 | 116 | 4.31 | 2.99 | 2.86, 3.29 | Yes | Yes | No |
| 42 | 0 | 107 | 0 | 2.83 | 2.74, 3.16 | No | No | Yes |
| 43 | 2 | 83 | 2.41 | 2.96 | 2.84, 3.27 | No | No | No |
| 44 | 0 | 98 | 0 | 2.99 | 2.87, 3.3 | No | No | Yes |

**Conclusions**

Overall, the majority of hospitals’ observed percentage died were outside the bootstrap CI confidence intervals, and the majority of their ratios of observed to expected were either higher than 1.2 or less than 0.8. Therefore, the large majority of hospitals (39/43, or 90.7%) were either unusually high or unusually low and require a site visit. Therefore, it may be worth considering changing the range in the future for which the observed to expected ratio is evaluated, since a large number of hospital site visits will be required.

This study is limited due to missing data situations throughout the data set. The large amount of missing data due to albumin was originally a concern, but once it was deemed non-valuable as a predictor, this missing data issue was no longer concerning.

There were smaller levels of missing data in other covariates that created limitations in this study. First, hospital 30 did not have any complete cases due to missing BMI values, so predicted mortality rates and bootstrapped confidence intervals could not be determined for this hospital in period 39.

However, the largest missing data concern is due to the bias that arises since those who are more likely to be missing covariate data had much higher rates of death than those who were not missing any covariate data. Those were not missing any covariate data had a death rate of 2.89% in period 39, while those who were missing covariate data had a death rate of 8.41%. Since overall missing covariate data is related to the outcome, this is concerning and likely has caused bias. This bias should be considered when interpreting these results, since the predicted values and resulting confidence intervals are likely too low.

**Code**

The code for this project can be found at <https://github.com/BIOS6623-UCD/bios6623-johnsra3/tree/master/Project2/Code>.

#==========================================================#

# Project 2 Data Cleaning

# Purpose: Clean data, save cleaned .csv file

# Rachel Johnson

#==========================================================#

#==========================================================#

# Import data and load libraries

#==========================================================#

library(haven)

setwd("~/School/AdvancedData")

vadata <- read\_sas("~/School/AdvancedData/vadata2.sas7bdat")

#==========================================================#

# Divide into different data sets (will use later)

#==========================================================#

rec <- vadata[vadata$sixmonth == 39, ]

old <- vadata[vadata$sixmonth != 39, ]

#==========================================================#

# Look at BMI issues in most recent period

#==========================================================#

plot((rec$weight)/(rec$height^2) \* 703, rec$bmi)

abline(1, 1)

hosp1\_5 <- rec[rec$hospcode >= 1 & rec$hospcode < 6, ]

plot((hosp1\_5$weight)/(hosp1\_5$height^2) \* 703, hosp1\_5$bmi)

abline(1, 1)

#All 5 of these hosp code BMIs are wrong (1 extraneous)

hosp6\_10 <- rec[rec$hospcode > 5 & rec$hospcode < 11, ]

plot((hosp6\_10$weight)/(hosp6\_10$height^2) \* 703, hosp6\_10$bmi)

abline(1, 1)

#All 5 of these hosp code BMIs are wrong (0 extraneous)

hosp11\_15 <- rec[rec$hospcode > 10 & rec$hospcode < 16, ]

plot((hosp11\_15$weight)/(hosp11\_15$height^2) \* 703, hosp11\_15$bmi)

abline(1, 1)

#All 5 of these hosp code BMIs are wrong (0 extraneous)

hosp16 <- rec[rec$hospcode == 16, ]

plot((hosp16$weight)/(hosp16$height^2) \* 703, hosp16$bmi)

abline(1, 1)

#Wrong BMI still :(

#==========================================================#

# Cleaning: fix BMI calculation issues

#==========================================================#

bmifix <- rec[rec$hospcode <= 16, ]

#All issues w/ weight and not height

bmifix$weight <- bmifix$weight \* 2.2

bmifix$bmi <- bmifix$weight/(bmifix$height^2) \* 703

plot((bmifix$weight)/(bmifix$height^2) \* 703, bmifix$bmi)

#One BMI is too low

#Hospital codes 17-44

hosp17plus <- rec[rec$hospcode > 16, ]

plot((hosp17plus$weight)/(hosp17plus$height^2) \* 703, hosp17plus$bmi)

abline(1, 1)

#All okay for calculation, 4 extraneous need to be fixed

hosp17plus$bmi\_check <- (hosp17plus$weight)/(hosp17plus$height^2) \* 703

hosp17plus$diff <- round(hosp17plus$bmi - hosp17plus$bmi\_check, 3)

plot(hosp17plus$bmi, hosp17plus$bmi\_check)

diffs <- hosp17plus[hosp17plus$diff != 0 & is.na(hosp17plus$diff) == F, ]

hosp17plus$bmi <- hosp17plus$bmi\_check

hosp17plus <- hosp17plus[, -c(which(colnames(hosp17plus) == "bmi\_check"),

which(colnames(hosp17plus) == "diff"))]

#Re-bind hospital code divided data

rec\_new <- rbind.data.frame(bmifix, hosp17plus)

plot((rec\_new$weight)/(rec\_new$height^2) \* 703, rec\_new$bmi)

abline(1, 1)

#All good now!

#Re-bind old and new data

vadata <- rbind.data.frame(rec\_new, old)

#==========================================================#

# Cleaning: remove proced = 2,

#==========================================================#

proc2 <- which(vadata$proced == 2)

vadata <- vadata[-proc2, ]

#==========================================================#

# Cleaning: collapse ASA categories

#==========================================================#

table(vadata$asa)

#some small categories--see if deaths in all groups

table(vadata$death30, vadata$asa)

#no deaths in asa=1; acc. to investigator, categ. 1/2/3 vs. 4/5

#note: 664 are missing

#Create indicator variable

vadata$asa\_indic <- ifelse(vadata$asa >= 4 & is.na(vadata$asa) == F, 1, 0)

#Collapse categories

vadata$asa <- ifelse(vadata$asa <= 3, "3 or less", vadata$asa)

vadata$asa <- ifelse(vadata$asa >=4, "4 or greater", vadata$asa)

vadata$asa <- factor(vadata$asa, levels = c("3 or less", "4 or greater"))

#==========================================================#

# Write csv of cleaned data

#==========================================================#

setwd("~/School/AdvancedData")

write.csv(vadata, "VadataCleaned.csv")

#==========================================================#

# Project 2 Table 1

# Purpose: Create Table 1

# Rachel Johnson

#==========================================================#

#==========================================================#

# Import data

#==========================================================#

setwd("~/School/AdvancedData")

vadata <- read.csv("VadataCleaned.csv", header = T)

#==========================================================#

# Divide into different data sets, create factor levels

#==========================================================#

rec <- vadata[vadata$sixmonth == 39, ]

old <- vadata[vadata$sixmonth != 39, ]

vadata$asa <- factor(vadata$asa, levels = c("3 or less", "4 or greater"))

#==========================================================#

# Make general table 1 (no stratifications)

#==========================================================#

tabdata <- vadata

tabdata <- tabdata[, -c(which(colnames(tabdata) == "weight"),

which(colnames(tabdata) == "height"),

which(colnames(tabdata) == "sixmonth"))]

names(tabdata)

rows <- 1 + 1 + 3 + 3 + 2 + 3 + 1

tab <- matrix(data = NA, nrow = rows, ncol = 2)

colnames(tab) <- c("", "")

tab[1, 1] <- "Hospitals (n)"

tab[1, 2] <- length(unique(tabdata$hospcode))

tab[2, 1] <- "Patients undergoing heart surgery (n)"

tab[2, 2] <- nrow(tabdata)

tab[3, 1] <- "Procedure (n (%))"

tab[4:5, 1] <- c("Valve surgery", "CABG surgery")

tab[4, 2] <- paste(nrow(tabdata[tabdata$proced == 0, ]), paste("(",

round(nrow(tabdata[tabdata$proced == 0, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[5, 2] <- paste(nrow(tabdata[tabdata$proced == 1, ]), paste("(",

round(nrow(tabdata[tabdata$proced == 1, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[6, 1] <- "ASA (n (%))"

tab[7:9, 1] <- c("3 or less", "4 or greater", "Missing")

tab[7, 2] <- paste(nrow(tabdata[tabdata$asa == "3 or less" & is.na(tabdata$asa) == F, ]), paste("(",

round(nrow(tabdata[tabdata$asa == "3 or less" & is.na(tabdata$asa) == F, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[8, 2] <- paste(nrow(tabdata[tabdata$asa == "4 or greater" & is.na(tabdata$asa) == F, ]), paste("(",

round(nrow(tabdata[tabdata$asa == "4 or greater" & is.na(tabdata$asa) == F, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[9, 2] <- paste(nrow(tabdata[is.na(tabdata$asa) == T, ]), paste("(",

round(nrow(tabdata[is.na(tabdata$asa) == T, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[10:11, 1] <- c("BMI (mean (SD))", "Missing (n (%))")

tab[10, 2] <- paste(round(mean(tabdata$bmi, na.rm = T), 2), paste("(",

round(sd(tabdata$bmi, na.rm = T), 2), ")", sep = ""))

tab[11, 2] <- paste(nrow(tabdata[is.na(tabdata$bmi) == T, ]), paste("(",

round(nrow(tabdata[is.na(tabdata$bmi) == T, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[12:13, 1] <- c("Albumin (mean (SD))", "Missing (n (%))")

tab[12, 2] <- paste(round(mean(tabdata$albumin, na.rm = T), 2), paste("(",

round(sd(tabdata$albumin, na.rm = T), 2), ")", sep = ""))

tab[13, 2] <- paste(nrow(tabdata[is.na(tabdata$albumin) == T, ]), paste("(",

round(nrow(tabdata[is.na(tabdata$albumin) == T, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

tab[14, 1] <- "30 day mortality (n (%))"

tab[14, 2] <- paste(nrow(tabdata[tabdata$death30 == 1 & is.na(tabdata$asa) == F, ]), paste("(",

round(nrow(tabdata[tabdata$death30 == 1 & is.na(tabdata$asa) == F, ])/nrow(tabdata) \* 100, 2), ")", sep = ""))

# setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

# write.csv(tab, "TableOverallCharacteristics.csv")

#==========================================================#

# Project 2 Missing Data

# Purpose: Explore missing data related to variable albumin

# Rachel Johnson

#==========================================================#

#==========================================================#

# Import data

#==========================================================#

library(dplyr)

setwd("~/School/AdvancedData")

vadata <- read.csv("VadataCleaned.csv", header = T)

#==========================================================#

# Look at missing data for ALBUMIN

#==========================================================#

#Look at covariates between missing and non-missing for ALBUMIN

alb\_miss <- vadata[is.na(vadata$albumin) == T, ]

alb\_pres <- vadata[is.na(vadata$albumin) == F, ]

table(alb\_miss$hospcode)

table(alb\_pres$hospcode)

#very similar

table(alb\_miss$proced)

table(alb\_pres$proced)

#very similar

table(alb\_miss$asa)

table(alb\_pres$asa)

#very similar

summary(alb\_miss$height)

summary(alb\_pres$height)

#very similar

summary(alb\_miss$weight)

summary(alb\_pres$weight)

#very similar

summary(alb\_miss$bmi) #some v. v. low measures, but are correct

summary(alb\_pres$bmi) #some v. v. low measures, but are correct

#very similar

table(alb\_miss$death30)

table(alb\_pres$death30)

#very similar

#No missingness is explained by covs, either MCAR or MNAR

#==========================================================#

# Compare outcome data for those w/ complete cases & those

# w/o complete cases for log. regr.

#==========================================================#

comp <- vadata[, c(1, 2, 3, 4, 8, 10, 11)]

comp <- comp[complete.cases(comp), ]

comp\_39 <- comp[comp$sixmonth == 39, ]

excl <- vadata[, c(1, 2, 3, 4, 8, 10, 11)]

excl <- excl[!excl$X %in% comp$X, ]

excl\_39 <- excl[excl$sixmonth == 39, ]

table(comp$death30)

table(excl$death30)

summary(comp$death30)

summary(excl$death30)

vadata$complete <- ifelse(vadata$X %in% comp$X, 1, 0)

pd39 <- vadata[vadata$sixmonth == 39, ]

chisq.test(pd39$complete, pd39$death30) #significantly different!--bias issue

#==========================================================#

# Project 2 Logistic Regressions

# Purpose: Run logistic regressions, create results tables

# Rachel Johnson

#==========================================================#

#==========================================================#

# Import data

#==========================================================#

library(boot)

setwd("~/School/AdvancedData")

vadata <- read.csv("VadataCleaned.csv", header = T)

#==========================================================#

# Divide into different data sets, create factor levels

#==========================================================#

rec <- vadata[vadata$sixmonth == 39, ]

old <- vadata[vadata$sixmonth != 39, ]

vadata$asa <- factor(vadata$asa, levels = c("3 or less", "4 or greater"))

#==========================================================#

# Logistic regressions

#==========================================================#

#With albumin

model1 <- glm(death30 ~ proced + asa + bmi + albumin, data = vadata,

family = binomial(link = "logit"))

summary(model1)

coeff1 <- summary(model1)$coefficients

#Without albumin

model2 <- glm(death30 ~ proced + asa + bmi, data = vadata,

family = binomial(link = "logit"))

summary(model2)

coeff2 <- summary(model2)$coefficients

#Decisions do not change-- albumin can be excluded, but still report results

#p-vals change some, but not in decision; estimates only change a little

#all estimate directions are the same

#albumin is not a significant predictor

#==========================================================#

# Make tables for logistic regression results

#==========================================================#

#Model 1- with albumin

mod1tab <- matrix(data = NA, nrow = 4, ncol = 4)

colnames(mod1tab) <- c("Covariate", "Estimate", "95% CI", "p-value")

mod1tab[1:4, 1] <- c("Procedure (reference = valve surgery)", "ASA (reference = 3 or less)", "BMI", "Albumin")

e <- exp(1)

mod1tab[1, 2] <- round(e^coeff1[2, 1], 3)

mod1tab[2, 2] <- round(e^coeff1[3, 1], 3)

mod1tab[3, 2] <- round(e^coeff1[4, 1], 3)

mod1tab[4, 2] <- round(e^coeff1[5, 1], 3)

mod1tab[1, 3] <- paste(paste(round(e^(coeff1[2, 1] - 1.96\*coeff1[2, 2]), 3), ",", sep = ""),

round(e^(coeff1[2, 1] + 1.96\*coeff1[2, 2]), 3))

mod1tab[2, 3] <- paste(paste(round(e^(coeff1[3, 1] - 1.96\*coeff1[3, 2]), 3), ",", sep = ""),

round(e^(coeff1[3, 1] + 1.96\*coeff1[3, 2]), 3))

mod1tab[3, 3] <- paste(paste(round(e^(coeff1[4, 1] - 1.96\*coeff1[4, 2]), 3), ",", sep = ""),

round(e^(coeff1[4, 1] + 1.96\*coeff1[4, 2]), 3))

mod1tab[4, 3] <- paste(paste(round(e^(coeff1[5, 1] - 1.96\*coeff1[5, 2]), 3), ",", sep = ""),

round(e^(coeff1[5, 1] + 1.96\*coeff1[5, 2]), 3))

mod1tab[1, 4] <- round(coeff1[2, 4], 3)

mod1tab[2, 4] <- round(coeff1[3, 4], 3)

mod1tab[2, 4] <- "<0.001"

mod1tab[3, 4] <- round(coeff1[4, 4], 3)

mod1tab[4, 4] <- round(coeff1[5, 4], 3)

#Model 2- without albumin

mod2tab <- matrix(data = NA, nrow = 3, ncol = 4)

colnames(mod2tab) <- c("Covariate", "Estimate", "95% CI", "p-value")

mod2tab[1:3, 1] <- c("Procedure (reference = valve surgery)", "ASA (reference = 3 or less)", "BMI")

e <- exp(1)

mod2tab[1, 2] <- round(e^coeff2[2, 1], 3)

mod2tab[2, 2] <- round(e^coeff2[3, 1], 3)

mod2tab[3, 2] <- round(e^coeff2[4, 1], 3)

mod2tab[1, 3] <- paste(paste(round(e^(coeff2[2, 1] - 1.96\*coeff2[2, 2]), 3), ",", sep = ""),

round(e^(coeff2[2, 1] + 1.96\*coeff2[2, 2]), 3))

mod2tab[2, 3] <- paste(paste(round(e^(coeff2[3, 1] - 1.96\*coeff2[3, 2]), 3), ",", sep = ""),

round(e^(coeff2[3, 1] + 1.96\*coeff2[3, 2]), 3))

mod2tab[3, 3] <- paste(paste(round(e^(coeff2[4, 1] - 1.96\*coeff2[4, 2]), 3), ",", sep = ""),

round(e^(coeff2[4, 1] + 1.96\*coeff2[4, 2]), 3))

mod2tab[1, 4] <- round(coeff2[2, 4], 3)

mod2tab[2, 4] <- round(coeff2[3, 4], 3)

mod2tab[2, 4] <- "<0.001"

mod2tab[3, 4] <- round(coeff2[4, 4], 3)

# setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

# write.csv(mod1tab, "ModelWithAlbuminResults.csv")

# write.csv(mod2tab, "ModelWithoutAlbuminResults.csv")

#==========================================================#

# Predicted values to individuals' odds to indivs' pred p

#==========================================================#

#Need a complete case data for BMI, procedure, ASA

comp <- vadata[, c(which(colnames(vadata) == "hospcode"),

which(colnames(vadata) == "sixmonth"),

which(colnames(vadata) == "proced"),

which(colnames(vadata) == "asa\_indic"),

which(colnames(vadata) == "bmi"),

which(colnames(vadata) == "death30"))]

comp <- comp[complete.cases(comp), ]

comp\_rec <- comp[comp$sixmonth == 39, ]

#Extract fitted values from summary of model2

(int <- summary(model2)$coefficients[1, 1])

(pred\_proced <- summary(model2)$coefficients[2, 1])

(pred\_asa <- summary(model2)$coefficients[3, 1])

(pred\_bmi <- summary(model2)$coefficients[4, 1])

#Create XB column in comp\_rec (last 6 month period only)

comp\_rec$XB <- int + pred\_proced\*comp\_rec$proced + pred\_asa\*comp\_rec$asa\_indic + pred\_bmi\*comp\_rec$bmi

#Create predicted p column in comp

comp\_rec$pred\_p <- inv.logit(comp\_rec$XB)

#==========================================================#

# Write csv of comp\_rec w/ predicted values by hosp

#==========================================================#

setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

write.csv(comp\_rec, "CompleteCasesPd39Predicted.csv")

#==========================================================#

# Project 2 Bootstrapping

# Rachel Johnson

#==========================================================#

#==========================================================#

# Import data

#==========================================================#

setwd("~/School/AdvancedData")

vadata <- read.csv("VadataCleaned.csv", header = T)

#==========================================================#

# Divide into different data sets, create factor levels

#==========================================================#

rec <- vadata[vadata$sixmonth == 39, ]

old <- vadata[vadata$sixmonth != 39, ]

vadata$asa <- factor(vadata$asa, levels = c("3 or less", "4 or greater"))

#==========================================================#

# Create data set for complete cases to use in bootstrap

#==========================================================#

comp <- vadata[, c(which(colnames(vadata) == "hospcode"),

which(colnames(vadata) == "sixmonth"),

which(colnames(vadata) == "proced"),

which(colnames(vadata) == "asa\_indic"),

which(colnames(vadata) == "bmi"),

which(colnames(vadata) == "death30"))]

comp <- comp[complete.cases(comp), ]

comp\_rec <- comp[comp$sixmonth == 39, ]

#==========================================================#

# Write loop to bootstrap

#==========================================================#

#Need following steps (1-5 in loop, 6 separate):

# 1. Sample from w/i total population w/ replacement in complete cases

# 2. Run logistic regression w/ this resampled population

# 3. Extract fitted values and exponentiate them

# 4. Calculate p\_fitted for each individual in the resampled data set w/ inv.logit

# 5. Get an average p\_fitted for each hospital w/ aggregate

# 6. Find distribution of p\_fitted for each hosp, w/ 2.5% and 97.5% pieces (boot.ci)

#Place to store p\_fits

num\_iter <- 10000

boot.stats <- matrix(data = NA, ncol = 43, nrow = num\_iter)

colnames(boot.stats) <- c(seq(from = 1, to = 29, by = 1), seq(from = 31, to = 44, by = 1))

for(i in 1:num\_iter){

set.seed(i)

boot.samps <- sample(nrow(comp), replace = T)

boot.dat <- comp[boot.samps, ]

boot.model <- glm(death30 ~ proced + asa\_indic + bmi, data = boot.dat, family = binomial(link = "logit"))

coeff <- summary(boot.model)$coefficients

xb <- coeff[1] + coeff[2]\*comp$proced + coeff[3]\*comp$asa\_indic + coeff[4]\*comp$bmi

comp$pfit <- inv.logit(xb)

comp\_pd39 <- comp[comp$sixmonth == 39, ]

boot.stats[i, ] <- round(aggregate(comp\_pd39$pfit, list(comp\_pd39$hospcode), mean)[, 2] \* 100, 2)

print(i)

}

setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

write.csv(boot.stats, "BootstrapResults\_raw.csv")

#==========================================================#

# Project 2

# Purpose: Create table of observed/expected/bootstrap res

# Rachel Johnson

#==========================================================#

#==========================================================#

# Import data

#==========================================================#

setwd("~/School/AdvancedData")

vadata <- read.csv("VadataCleaned.csv", header = T)

setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

comp\_rec <- read.csv("CompleteCasesPd39Predicted.csv", header = T)

boot.res <- read.csv("BootstrapResults\_raw.csv", header = T)

boot.res <- boot.res[, -1]

#==========================================================#

# Divide into different data sets, create factor levels

#==========================================================#

rec <- vadata[vadata$sixmonth == 39, ]

old <- vadata[vadata$sixmonth != 39, ]

vadata$asa <- factor(vadata$asa, levels = c("3 or less", "4 or greater"))

#==========================================================#

# Get confidence intervals from raw bootstrap data

#==========================================================#

boot.ci\_low <- round(apply(boot.res, 2, quantile, probs = 0.025), 2)

boot.ci\_high <- round(apply(boot.res, 2, quantile, probs = 0.975), 2)

boot.ci <- paste(paste(boot.ci\_low, ",", sep = ""), boot.ci\_high)

hosps <- c(seq(from = 1, to = 29, by = 1), seq(from = 31, to = 44, by = 1))

boottab <- cbind.data.frame(hosps, boot.ci)

colnames(boottab) <- c("Hospital", "95% CI")

#==========================================================#

# Create table 2 w/ hospital death rates

#==========================================================#

tab2 <- matrix(data = NA, nrow = length(unique(rec$hospcode)), ncol = 5)

colnames(tab2) <- c("Hospital", "Patients died", "Patients in surgery", "Percent died",

"Predicted percent died for last 6 months (population-adjusted)")

tab2[1:44, 1] <- unique(rec$hospcode)

tab2[1:44, 2] <- aggregate(rec$death30, list(rec$hospcode), sum)[, 2]

tab2[1:44, 3] <- aggregate(rec$death30, list(rec$hospcode), length)[, 2]

tab2[1:44, 4] <- round(tab2[, 2]/tab2[, 3] \* 100, 2)

#Note: hospital 30 does not have any data for last period, so it will not be estimated for last pd.

hosp1\_29 <- as.list(unique(comp\_rec$hospcode)[1:29])

p1\_29 <- as.data.frame(comp\_rec[comp\_rec$hospcode %in% hosp1\_29, c(which(colnames(comp\_rec) == "pred\_p"),

which(colnames(comp\_rec) == "hospcode"))])

tab2[1:29, 5] <- round(aggregate(p1\_29$pred\_p, list(p1\_29$hospcode), mean)[, 2] \* 100, 2)

hosp31\_44 <- seq(from = 31, to = 44, by = 1)

p31\_44 <- as.data.frame(comp\_rec[comp\_rec$hospcode %in% hosp31\_44, c(which(colnames(comp\_rec) == "pred\_p"),

which(colnames(comp\_rec) == "hospcode"))])

tab2[31:44, 5] <- round(aggregate(p31\_44$pred\_p, list(p31\_44$hospcode), mean)[, 2] \* 100, 2)

# setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

# write.csv(tab2, "TableDeathsByHospital.csv")

#==========================================================#

# Update hospital table w/ bootstrap CIs

#==========================================================#

tab2\_first <- tab2[1:29, ]

tab2\_30 <- as.data.frame(tab2[30, ])

tab2\_30 <- t(tab2\_30)

tab2\_end <- tab2[31:44, ]

boottab\_first <- boottab[1:29, ]

boottab\_end <- boottab[30:43, ]

finaltab\_first <- merge(tab2\_first, boottab\_first, by = "Hospital")

finaltab\_30 <- matrix(data = NA, nrow = 1, ncol = 6)

colnames(finaltab\_30) <- colnames(finaltab\_first)

finaltab\_30[,1:5] <- tab2\_30[, 1:5]

finaltab\_end <- merge(tab2\_end, boottab\_end, by = "Hospital")

finaltab <- rbind.data.frame(finaltab\_first, finaltab\_30, finaltab\_end)

#==========================================================#

# Hosp observed p/expected p > 1.2

#==========================================================#

finaltab <- as.data.frame(finaltab)

finaltab$`Observed Above Confidence Interval?` <- ifelse(finaltab$`Percent died` > boot.ci\_high,

"Yes", "No")

finaltab$`Observed Above Confidence Interval?`[30] <- NA

#==========================================================#

# Hosp observed p/expected p > 1.2

#==========================================================#

finaltab$`Unusually High?` <- ifelse(finaltab$`Percent died`/finaltab$`Predicted percent died for last 6 months (population-adjusted)` > 1.2,

"Yes", "No")

table(finaltab$`Unusually High?`)

#17 hospitals were unusually high

#==========================================================#

# Hosp observed p/expected p < 0.8

#==========================================================#

finaltab$`Unusually Low?` <- ifelse(finaltab$`Percent died`/finaltab$`Predicted percent died for last 6 months (population-adjusted)` < 0.8,

"Yes", "No")

table(finaltab$`Unusually Low?`)

# setwd("C:/Repositories/bios6623-johnsra3/Project2/Reports")

# write.csv(finaltab, "TableExpectedPropsBootstrap.csv")