Stat 472 Homework 4 - Calculus Retention

Samantha Bothwell, Crystal Wu, and Wulf Novak

April 2, 2019

Mixed Effects Logistic Regression Model

$$
\begin{aligned}
log\left(\frac{p_i}{1-p_i}\right) =& \beta_0 + \beta_1 \cdot 1[\text{CollegeCalc}_i] + \beta_2 \cdot 1[\text{NoCalc}_i] + \beta_3 \cdot 1[\text{Engineering}_i] \\
&+ \beta_4 \cdot 1[\text{Pre-med}_i] + \beta_5 \cdot 1[\text{Non-STEM}_i] + \beta_6 \cdot 1[\text{Undecided}_i] \\
&+ \beta_7 \cdot \text{StdTest}_i + \beta_8 \cdot \text{InstructorQuality}_i + \beta_9 \cdot \text{StudentPractices}_i \\
&+ \beta_1 0 \cdot 1[\text{Female}_i] + \alpha_{\text{Institution}_i}
\end{aligned}
$$

Where $P(Y_i = 1) = p_i$ and $\alpha_{\text{Institution}_i} \sim N(0, \sigma^2)$

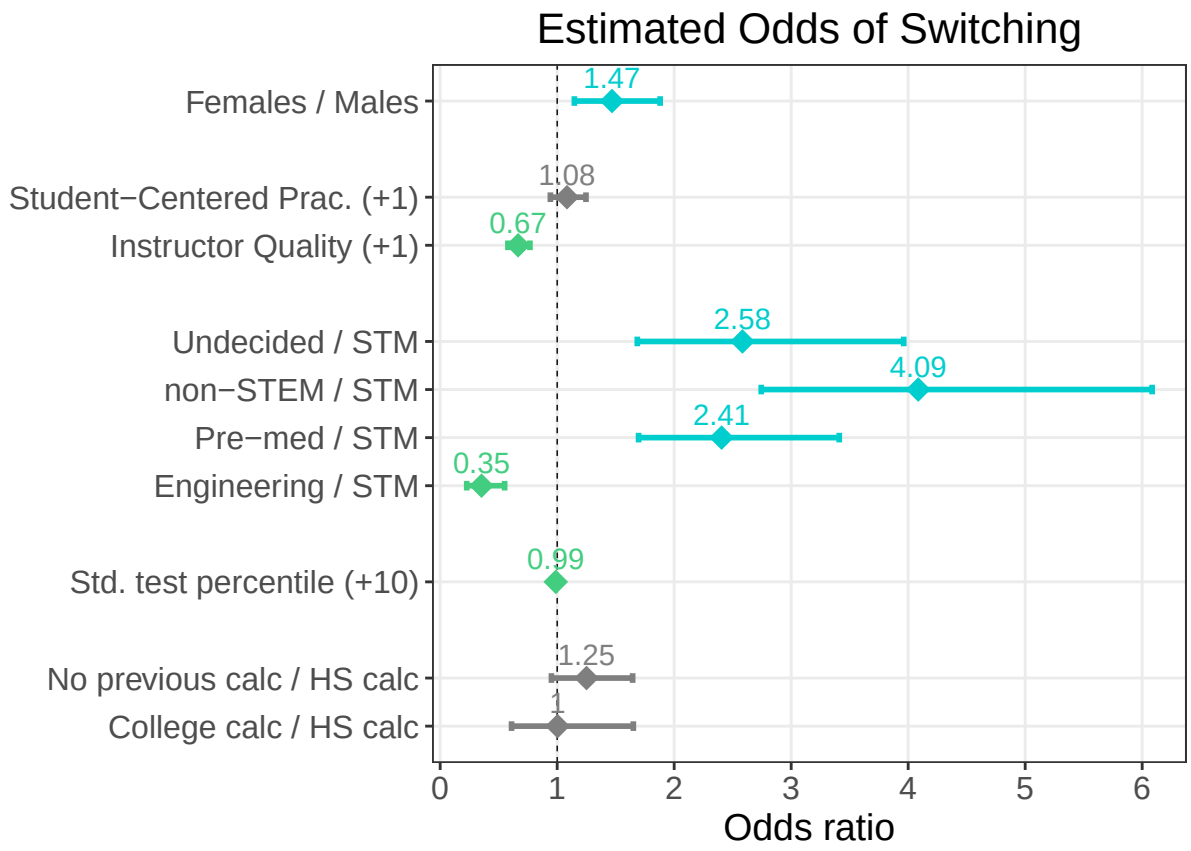
The categories high school calculus experience, STM career, and male were specified as the baseline categories (set equal to 0). The indicator function 1[A] is defined as 1 if A is true and 0 if A if false. The β coefficients corresponding to the categorical variables are interpreted as the change in the odds of switching for each category compared to the respective baseline category. The Institution variable is treated as a random effect.

Odds Ratio Table

	Odds Ratio	Odds Ratio CI
Previous College Calculus	1.003	(1.147 , 1.88)
No Previous Calculus	1.251	(0.942 , 1.246)
Career Choice: Engineering	0.353	(0.579 , 0.767)
Career Choice: Premed	2.405	(1.685 , 3.962)
Career Choice: Non-STEM	4.086	(2.744 , 6.085)
Career Choice: Undecided	2.584	(1.696 , 3.411)
Percentile (10 pt increase)	0.988	(0.227 , 0.551)
Instructor Quality (1 pt increase)	0.666	(0.979 , 0.997)
Student-Centered Practices (1 pt increase)	1.084	(0.951 , 1.645)
Women	1.469	(0.61 , 1.651)

We used this table to compare our results to table 9 in the paper. Our estimates do differ slightly since we used a different source for SAT and ACT percentiles. Another reason for the different estimates could be that the orginal paper used a Bayesian model but we used a generalized linear model.

Fig 2. Odds ratios of switching for student attributes

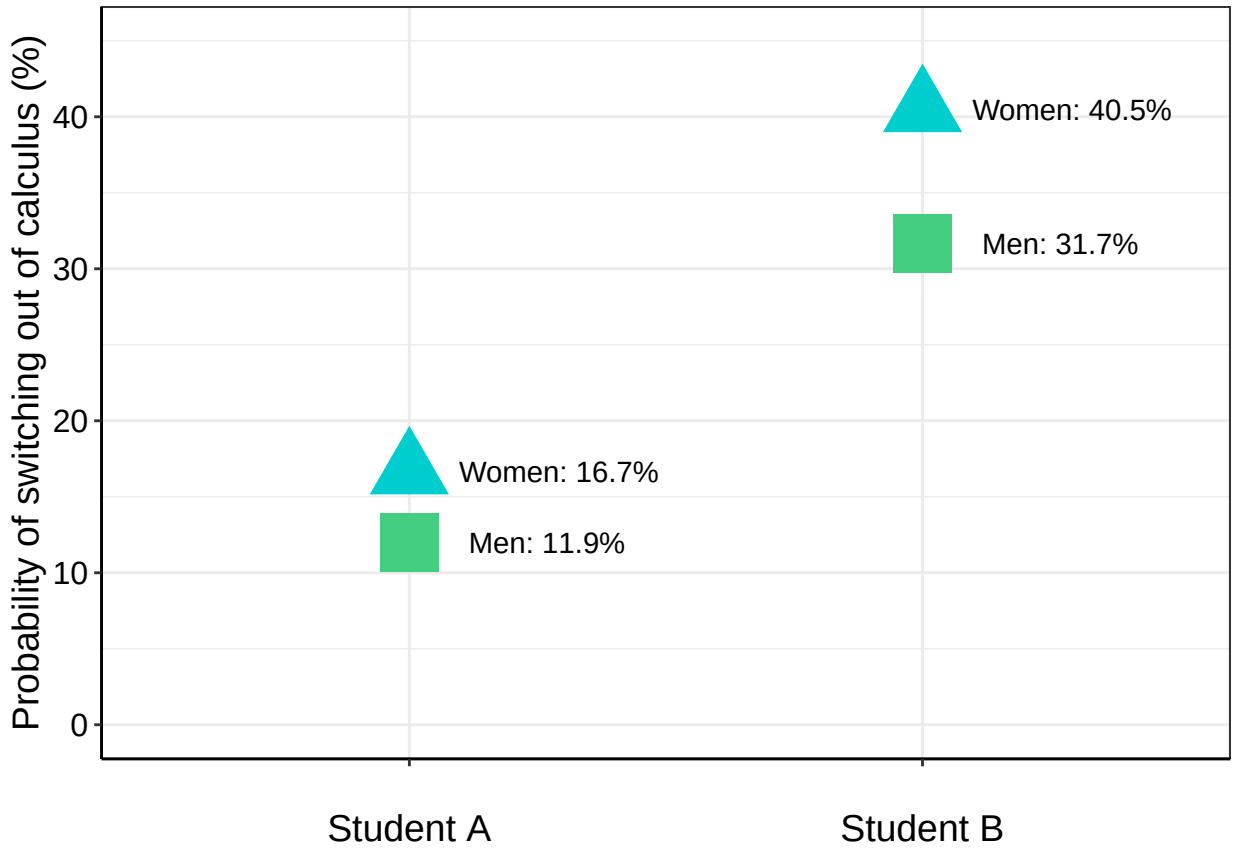


The diamond represents the odds ratio estimate and the bars represent the 95% credible interval. The continuous variables noted with (+x) on the left compare a student who reported x-points higher than another student. Labels of the form A/B correspond to the ratio of the odds of switching for a student of type A to the odds of switching for a student of type B. Variables associated with decreased likeliness and increased likeliness of switching are highlighted in green and teal, respectively.[N=2266.]

Analysis of variables:

- **Females / Males:** Females are 1.47 times as likely to switch out of calculus as men.
- **Student-Centered Prac. (+1):** A student who reports +1 point higher than another student on Student-Centered Practices will be 1.08 times as likely to switch out of calculus. (Higher rating on Student-Centered Practices → more likely to switch out of calculus)
- **Instructor Quality (+1):** A student who reports +1 point higher than another student on Instructor Quality will be 0.67 times as likely to switch out of calculus. (Higher rating on Instructor Quality → less likely to switch out of calculus)
- **Undecided / STM:** A student who indicated their career choice as "Undecided" is 2.58 times as likely to switch out of calculus as a student who indicated their career choice as a STM field.
- **non-STEM / STM:** A student who indicated their career choice as "non-STEM" is 4.09 times as likely to switch out of calculus as a student who indicated their career choice as a STM field.
- **Pre-med / STM:** A student who indicated their career choice as "Pre-med" is 2.41 times as likely to switch out of calculus as a student who indicated their career choice as a STM field.
- **Engineering / STM:** A student who indicated their career choice as "Engineering" is 2.58 times as likely to switch out of calculus as a student who indicated their career choice as a STM field.
- **Std. test percentile (+10):** A student who scored +10 percentile points higher than another student on the SAT/ACT will be 0.99 times as likely to switch out of calculus.
- **No previous calc / HS calc:** A student who had no previous calculus experience is 1.25 times as likely to switch out of calculus as a student who had high school calculus.
- **College calc / HS calc:** A student who had college calculus experience is equally as likely to switch out of calculus as a student who had high school calculus.

Fig 3. Comparison of probability of switching for two hypothetical students



- Average standardized math score
- Took high school calculus
- Pursuing STM
- Reports average levels of Student-Centered Practices and lower than average levels of Instructor Quality

- Average standardized math score
- No high school calculus
- Pursuing non-STM
- Reports average levels of Student-Centered Practices and Instructor Quality

Note: We used the average of Instructor Quality - 1 to represent lower than average levels of Instructor Quality for Student A.

Appendix

```
# Data Preparation
S1dat <- read.csv("D:/School/Spring 2019/Stat 472/Calculus Retention/FullData.CSV")
S1dat$FUS <- paste(S1dat$Q3FUS_No,S1dat$Q3FUS_Yes)

S1dat$FUS[S1dat$FUS=="   "] <- "NA"; S1dat$FUS[S1dat$FUS == "No  "] <- "No"
S1dat$FUS[S1dat$FUS=="  Yes"] <- "Yes"
S1dat$Q26[is.na(S1dat$Q26)] <- 0
S1dat$Q3Post[is.na(S1dat$Q3Post)] <- 0
S1dat$Q5Post[is.na(S1dat$Q5Post)] <- 0
```

```
# Switch Persist

S1dat$SwitchPersist <- ifelse(S1dat$Q26 == 1 & S1dat$FUS == "No", 1,
                       ifelse(S1dat$Q26 == 3 & S1dat$FUS == "No", 2,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 1 & S1dat$FUS == "No", 3,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 3 & S1dat$FUS == "No", 4,
                       ifelse(S1dat$Q26 == 1 & S1dat$Q5Post == 1 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 5,
                       ifelse(S1dat$Q26 == 1 & S1dat$Q3Post == 2 & S1dat$FUS == "NA", 6,
                       ifelse(S1dat$Q26 == 3 & S1dat$Q5Post == 1 & S1dat$FUS == "NA" &
                                S1dat$Q3Post == 3, 7,
                       ifelse(S1dat$Q26 == 3 & S1dat$Q3Post == 2 & S1dat$FUS == "NA", 8,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 1 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 9,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 1 & S1dat$Q3Post == 2 &
                                S1dat$FUS == "NA", 10,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 3 & S1dat$Q3Post == 2 &
                                S1dat$FUS == "NA", 11,
                       ifelse(S1dat$Q26 == 1 & S1dat$FUS == "Yes", 12,
                       ifelse(S1dat$Q26 == 3 & S1dat$FUS == "Yes", 13,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 1 & S1dat$FUS == "Yes", 14,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 3 & S1dat$FUS == "Yes", 15,
                       ifelse(S1dat$Q26 == 1 & S1dat$Q3Post == 1 & S1dat$FUS == "NA", 16,
                       ifelse(S1dat$Q26 == 1 & S1dat$Q5Post == 3 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 17,
                       ifelse(S1dat$Q26 == 1 & S1dat$Q5Post == 2 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 18,
                       ifelse(S1dat$Q26 == 1 & S1dat$Q5Post == 0 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 19,
                       ifelse(S1dat$Q26 == 3 & S1dat$Q3Post == 1 & S1dat$FUS == "NA", 20,
                       ifelse(S1dat$Q26 == 3 & S1dat$Q5Post == 3 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 21,
                       ifelse(S1dat$Q26 == 3 & S1dat$Q5Post == 2 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 22,
                       ifelse(S1dat$Q26 == 3 & S1dat$Q5Post == 0 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 23,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 1 & S1dat$Q3Post == 1 &
                                S1dat$FUS == "NA", 24,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 3 & S1dat$Q3Post == 1 &
                                S1dat$FUS == "NA", 25,
                       ifelse(S1dat$Q26 == 0 & S1dat$Q5Post == 3 & S1dat$Q3Post == 3 &
                                S1dat$FUS == "NA", 26,
                              0))))))))))))))))))))))))))
S1dat$SP <- ifelse(S1dat$SwitchPersist == 0, "NA",
            ifelse(S1dat$SwitchPersist <= 11, "S", "P"))
```

```
S1dat$Q18Post_Discouraged <- 7 - S1dat$Q18Post_Discouraged
instructor <- subset(S1dat, S1dat$Q18Post_Applications > 0 &
                      S1dat$Q18Post_Appointments > 0 &
                      S1dat$Q18Post_AskedQs > 0 &
                      S1dat$Q18Post_Discouraged > 0 &
                      S1dat$Q18Post_Explanations > 0 &
                      S1dat$Q18Post_Listened > 0 &
                      S1dat$Q18Post_ProblemSolver > 0 &
                      S1dat$Q18Post_Time > 0)
instructor <- instructor[,c(22:29)]
p1 <- prcomp(na.omit(instructor))
```

```
pca <- p1$rotation[,1]
pca_scaled <- rep(NA,length(pca))
for(i in 1:length(pca)){
  pca_scaled[i] <- pca[i]/sum(pca)
}
S1dat$IQ <- apply(S1dat[22:29], 1, weighted.mean, w=pca_scaled, na.rm=TRUE)

S1dat$Q19Post_Lecture <- 7-S1dat$Q19Post_Lecture
instructor_19 <- subset(S1dat, !(is.na(S1dat$Q19Post_SpecificProblems) | is.na(S1dat$Q19Post_WorkTogether
                        | is.na(S1dat$Q19Post_Discussion) | is.na(S1dat$Q19Post_Presentations) |
                          is.na(S1dat$Q19Post_Individually) | is.na(S1dat$Q19Post_Lecture) |
                          is.na(S1dat$Q19Post_AskQuestions) | is.na(S1dat$Q19Post_ExplainThinking))))
instructor_19 <- instructor_19[,c(30:37)]
p2 <- prcomp(na.omit(instructor_19))
pca1 <- round(p2$rotation[,1],3)
pca1_scaled <- rep(NA,length(pca1))
for(i in 1:length(pca1)){
  pca1_scaled[i] <- pca1[i]/sum(pca1)
}
S1dat$SCP <- apply(S1dat[30:37], 1, weighted.mean, w=pca1_scaled, na.rm=TRUE)
```

```
# Clean data
switchers <- subset(S1dat, SwitchPersist != 0)

cleandat <- switchers[complete.cases(switchers$Q48 & switchers$Q60 & (switchers$Q3_SATMath |
                     switchers$Q7_ACTMath) & (switchers$Q18Post_AskedQs | switchers$Q18Post_Listened |
                     switchers$Q18Post_Applications | switchers$Q18Post_Appointments) &
                     (switchers$Q19Post_AskQuestions | switchers$Q19Post_SpecificProblems |
                     switchers$Q19Post_WorkTogether | switchers$Q19Post_Discussion |
                     switchers$Q19Post_Presentations | switchers$Q19Post_Individually |
                     switchers$Q19Post_Lecture | switchers$Q19Post_ExplainThinking) & (switchers$Q18 |
                     switchers$Q15_CalculusNonAPFinalGrade | switchers$Q17_CalculusABFinalGrade |
                     switchers$Q17_CalculusBCFinalGrade)),]

cleandata <- cleandat[which((cleandat$Q3_SATMath <= 800 & cleandat$Q3_SATMath >= 200 &
                     cleandat$Q3_SATMath%%10 == "0") |
                     (cleandat$Q7_ACTMath <= 36 & cleandat$Q7_ACTMath >= 1)),]
```

```
# Percentiles

SATscore <- rev(seq(200,800,by=10)) #61 numbers
SATper <- c(99,99,99,98,97,97,96,95,95,94,93,91,90,88,87,85,83,82,79,77,75,73,70,67,64,62,59,55,52,49,45,
            42,40,36,33,30,27,24,21,19,16,14,12,10,9,7,6,5,4,3,3,2,2,1,1,1,1,1,0,0,0)

SATPerc = SATper[match(cleandata$Q3_SATMath,SATscore)]
cleandata <- cbind(cleandata, SATPerc)

ACTscore <- 36:1
ACTper <- c(99,99,99,99,98,96,95,92,90,87,83,79,74,68,62,56,50,43,36,30,24,18,12,7,4,1,1,1,1,1,1,1,1,1,1,1,1)

ACTPerc = ACTper[match(cleandata$Q7_ACTMath, ACTscore)]
cleandata <- cbind(cleandata, ACTPerc)

#creating average standardized test percentile
Perc <- rowMeans(cbind(SATPerc,ACTPerc), na.rm = T )
cleandata <- cbind(cleandata, Perc)

cleandata$ACTPerc <- ifelse(cleandata$SATPerc >= 1 & is.na(cleandata$ACTPerc),
                        cleandata$SATPerc, cleandata$ACTPerc)
cleandata$SATPerc <- ifelse(cleandata$ACTPerc >= 1 & is.na(cleandata$SATPerc),
                        cleandata$ACTPerc, cleandata$SATPerc)
cleandata$Perc <- ifelse(cleandata$ACTPerc >= 1 & cleandata$SATPerc >= 1,
                     (cleandata$ACTPerc + cleandata$SATPerc)/2, "")
```

```
cleandata$SP[cleandata$SP == "S"] = 1
cleandata$SP[cleandata$SP == "P"] = 0
```

```
# Calc Experience
```

```r
X_prev_calc<-rep(NA,nrow(cleandata))
X_prev_calc[cleandata$Q15_CalculusNonAPFinalGrade>0]<-"high"
X_prev_calc[cleandata$Q17_CalculusABFinalGrade>0]<-"high"
X_prev_calc[cleandata$Q17_CalculusBCFinalGrade>0]<-"high"
X_prev_calc[cleandata$Q18 == 1]<-"college"
X_prev_calc[is.na(X_prev_calc) & cleandata$Q18 == 2]<-"none"

cleandata <- cbind(cleandata, X_prev_calc)
```

```r
# Indicator Variables

cleandata$CollegeCalc <- ifelse(cleandata$X_prev_calc == "college", 1, 0)
cleandata$NoCalc <- ifelse(cleandata$X_prev_calc == "none", 1, 0)
```

```r
cleandata$Q48[cleandata$Q48 == 1] = 0 # Male
cleandata$Q48[cleandata$Q48 == 2] = 1 # Female
```

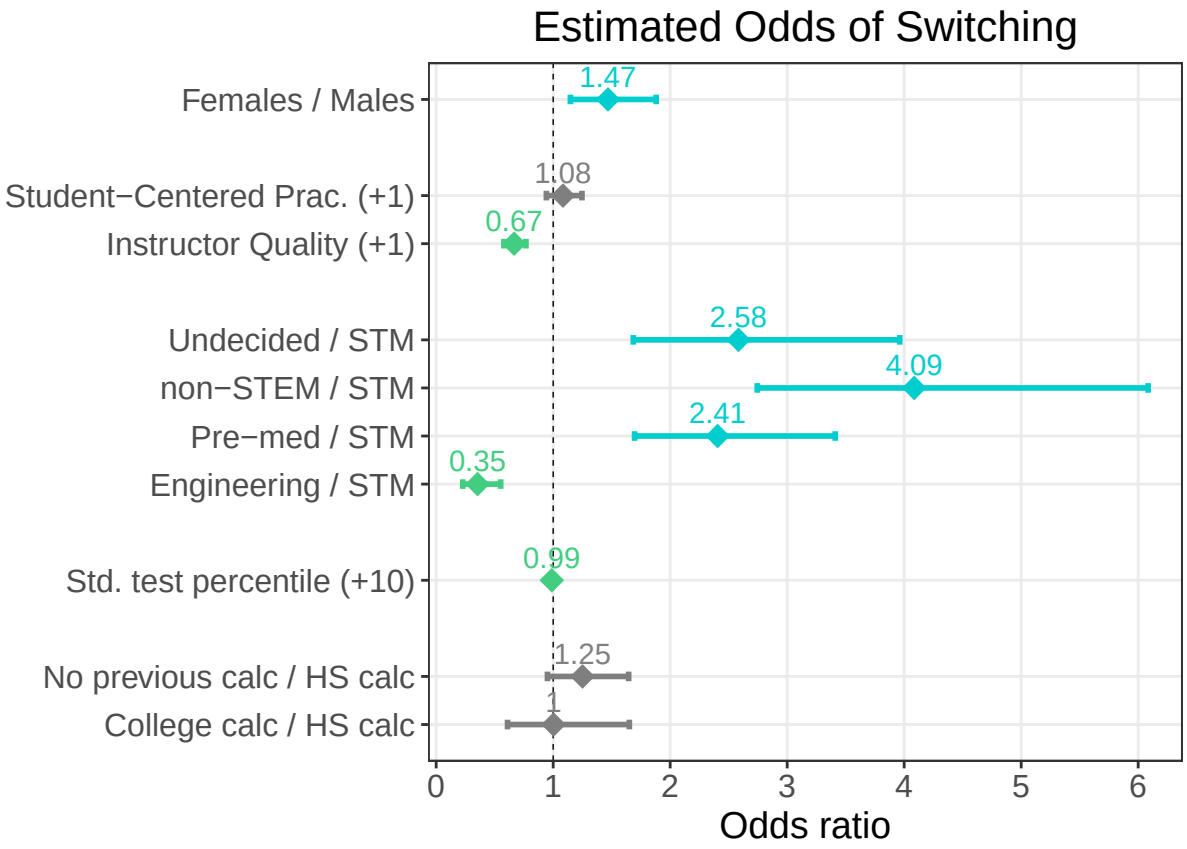
```r
X_career<-rep(NA,nrow(cleandata))
X_career[cleandata$Q60%in%c(3,4,5,7,8,9)]<-"STM"
X_career[cleandata$Q60 == 6]<-"Engineering"
X_career[cleandata$Q60%in%c(1,2)]<-"Pre-med"
X_career[cleandata$Q60%in%c(10,11,12,13,14,15)]<-"Non-STEM"
X_career[cleandata$Q60 == 16]<-"Undecided"

cleandata <- cbind(cleandata, X_career)

cleandata$engineering <- ifelse(cleandata$X_career == "Engineering",1,0)
cleandata$stem <- ifelse(cleandata$X_career == "STM",1,0)
cleandata$premed <- ifelse(cleandata$X_career == "Pre-med",1,0)
cleandata$nonstem <- ifelse(cleandata$X_career == "Non-STEM",1,0)
cleandata$undecided <- ifelse(cleandata$X_career == "Undecided",1,0)
```

```r
library(lme4)
options(scipen=5)
cleandata$SP <- as.numeric(cleandata$SP)
model <- glmer(SP~CollegeCalc+NoCalc+engineering+premed+nonstem+undecided+Perc+IQ+SCP+
                 Q48+(1|Institution),
               data = cleandata, family="binomial")
sum <- summary(model)
```

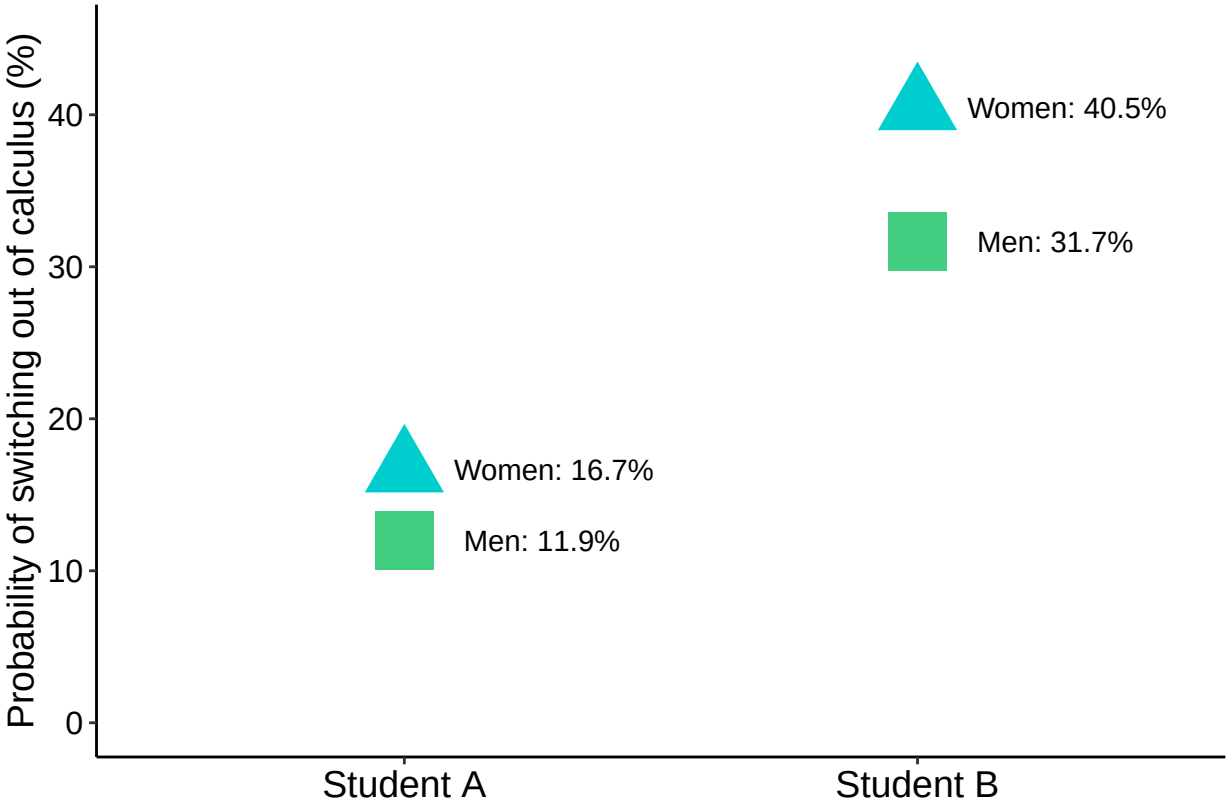
```r
or <- exp(sum$coefficients)
sum_est <- sum$coefficients[c(11,10,9,7,6,5,4,8,3,2)]
sum_se <- sum$coefficients[c(11,10,9,7,6,5,4,8,3,2),2]
lower <- sum_est - 1.96*sum_se
upper <- sum_est + 1.96*sum_se
or_est <- exp(sum_est)
or_lower <- round(exp(lower),3); or_upper <- round(exp(upper),3)
CI <- rbind(or_lower, or_upper)

ORc1 <- round(exp(sum$coefficients[2:11,1]),3)
ORc2 <- paste("(",or_lower,",",or_upper,")")
ORdt <- cbind(ORc1, ORc2)
colnames(ORdt) <- c("Odds Ratio", "Odds Ratio CI")
rownames(ORdt) <- c("Previous College Calculus", "No Previous Calculus",
                    "Career Choice: Engineering", "Career Choice: Premed",
                    "Career Choice: Non-STEM", "Career Choice: Undecided",
                    "Percentile (10 pt increase)",
                    "Instructor Quality (1 pt increase)",
                    "Student-Centered Practices (1 pt increase)", "Women")

library(kableExtra)
kable(ORdt, "latex", booktabs = T, linesep = "", align = "c") %>%
  kable_styling(full_width = F, latex_options = "striped") %>%
  column_spec(1, width = "50mm", border_right = T) %>%
  column_spec(2, width = "25mm", border_right = T)
```

```r
# Figure 2
library(ggplot2)
```

```r
# Create labels
boxLabels = c("Females / Males", "Student-Centered Prac. (+1)", "Instructor Quality (+1)",
              "Undecided / STM", "non-STEM / STM", "Pre-med / STM", "Engineering / STM",
              "Std. test percentile (+10)", "No previous calc / HS calc", "College calc / HS calc")

# Enter summary data. boxOdds are the odds ratios, boxCILow is the lower bound of the CI,
# boxCIHigh is the upper bound.
yAxis = c(14,12,11,9,8,7,6,4,2,1)
df <- data.frame(
  boxOdds = or_est,
  boxCILow = or_lower,
  boxCIHigh = or_upper
)

# Plot
p <- ggplot(df, aes(x = boxOdds, y = yAxis))
p + geom_vline(aes(xintercept = 1), size = .25, linetype = "dashed") +
  geom_errorbarh(aes(xmax = boxCIHigh, xmin = boxCILow), size = 1, height = .2,
                 color = ifelse(or_upper < 1, "seagreen3", ifelse(or_lower > 1, "cyan3", "grey50"))) +
  geom_point(size = 4, pch = 18,
             color = ifelse(or_upper < 1, "seagreen3", ifelse(or_lower > 1, "cyan3", "grey50"))) +
  theme_bw() +
  theme(panel.grid.minor = element_blank()) +
  scale_y_continuous(breaks = yAxis, labels = boxLabels) +
  scale_x_continuous(breaks = seq(0,7,1) ) +
  ylab("") +
  xlab("Odds ratio") +
  ggtitle("Estimated Odds of Switching") +
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        axis.text=element_text(size=12),
        axis.title=element_text(size=14)) +
  geom_text(aes(label = round(boxOdds,2)),vjust = -0.6,
            color = ifelse(or_upper < 1, "seagreen3", ifelse(or_lower > 1, "cyan3", "grey50")))
```




```
stdA_female <- t( c(0,0,0,0,0,0,mean(cleandata$Perc),mean(cleandata$IQ) - 1,mean(cleandata$SCP),1,0))
stdA_male <- t( c(0,0,0,0,0,0,mean(cleandata$Perc),mean(cleandata$IQ) - 1,mean(cleandata$SCP),0,0))
stdB_female <- t( c(0,1,0,0,1,0,mean(cleandata$Perc),mean(cleandata$IQ),mean(cleandata$SCP),1,0))
stdB_male <- t( c(0,1,0,0,1,0,mean(cleandata$Perc),mean(cleandata$IQ),mean(cleandata$SCP),0,0))
newdata <- as.data.frame(rbind(stdA_female, stdB_female, stdA_male, stdB_male))
colnames(newdata) <- c("CollegeCalc", "NoCalc", "engineering", "premed", "nonstem", "undecided",
                       "Perc", "IQ", "SCP", "Q48", "Institution")
pred <- 100*predict(model, newdata = newdata, re.form=NA, type = "response")
Student <- c("Student A", "Student B", "Student A", "Student B")
Gender <- c("Female", "Female", "Male", "Male")
Labels <- c("Women: 16.7%", "Women: 40.5%", "  Men: 11.9%", "  Men: 31.7%")
pred <- as.data.frame(cbind(pred, Student, Gender, Labels))
pred$pred <- as.numeric(as.character(pred$pred))

q <-ggplot(pred, aes(x = Student, y = pred)) +
  scale_y_continuous(breaks = seq(0,45,10), limit=c(0,45)) +
  geom_point(pch = ifelse(pred$Gender=="Female", 17, 15),
             color = ifelse(pred$Gender=="Female", "cyan3", "seagreen3"),
             size = 10) +
  geom_text(aes(label = Labels), hjust= -.25) +
  xlab("") +
  ylab("Probability of switching out of calculus (%)") +
  theme(axis.text=element_text(size=12)) + theme(panel.background = element_blank(),
                                                 axis.line = element_line(colour = "black"),
                                                 axis.text.y = element_text(color = '#000000', size=12),
                                                 axis.text.x = element_text( color = '#000000', size=14),
                                                 axis.title.y = element_text(colour = '#000000', size= 14))
q
```



- Average standardized math score
- Took high school calculus
- Pursuing STM
- Reports average levels of Student-Centered Practices and lower than average levels of Instructor Quality

- Average standardized math score
- No high school calculus
- Pursuing non-STM
- Reports average levels of Student-Centered Practices and Instructor Quality