Problem Set 3

Reed Sorensen

November 3, 2016

Group members: Orvalho Augusto, Max Griswold, Gloria Ikilezi, Chris Kemp

Question 1a

```
require(dplyr)
require(tile)
require(simcf)
require(RColorBrewer)
require(MASS)
require(ggplot2)

set.seed(123)

df <- read.csv("http://faculty.washington.edu/cadolph/mle/cyyoung.csv")

y <- df$cy
x <- cbind(df$era, df$winpct)

llk.logit <- function(param,y,x) {
  os <- rep(1,length(x[,1]))
  x <- cbind(os,x)
  b <- param[1:ncol(x)]
  xb <- x %*% b
  sum(-1 * (y*log(1+exp(-xb)) + (1-y)*log(1+exp(xb))))
}

stval <- lm(y~x)$coefficients

result1 <- optim(
  par = stval, fn = llk.logit,
  method = "BFGS", hessian = T, control = list(fnscale = -1),
  y = y, x = x
)


# optim results

matrix(
  c(round(result1$par, digits = 4), round(sqrt(diag(solve(-1 * result1$hessian))), digits = 4)),
  dimnames = list(c("(Intercept)", "era", "winpct"), c("Estimate", "Std. Error")),
  ncol = 2
)
```

```
##             Estimate Std. Error
## (Intercept)   1.3423     3.2899
```

```
## era          -2.1122     0.5130
## winpct        6.1798     3.9119
```

```
cat(paste("Log likelihood at maximum:", round(result1$value, digits = 3)))
```

```
## Log likelihood at maximum: -46.224
```

```
# glm results

fit1 <- glm(cy ~ era + winpct, family = "binomial", data = df)
summary(fit1)$coefficients[,1:2]
```

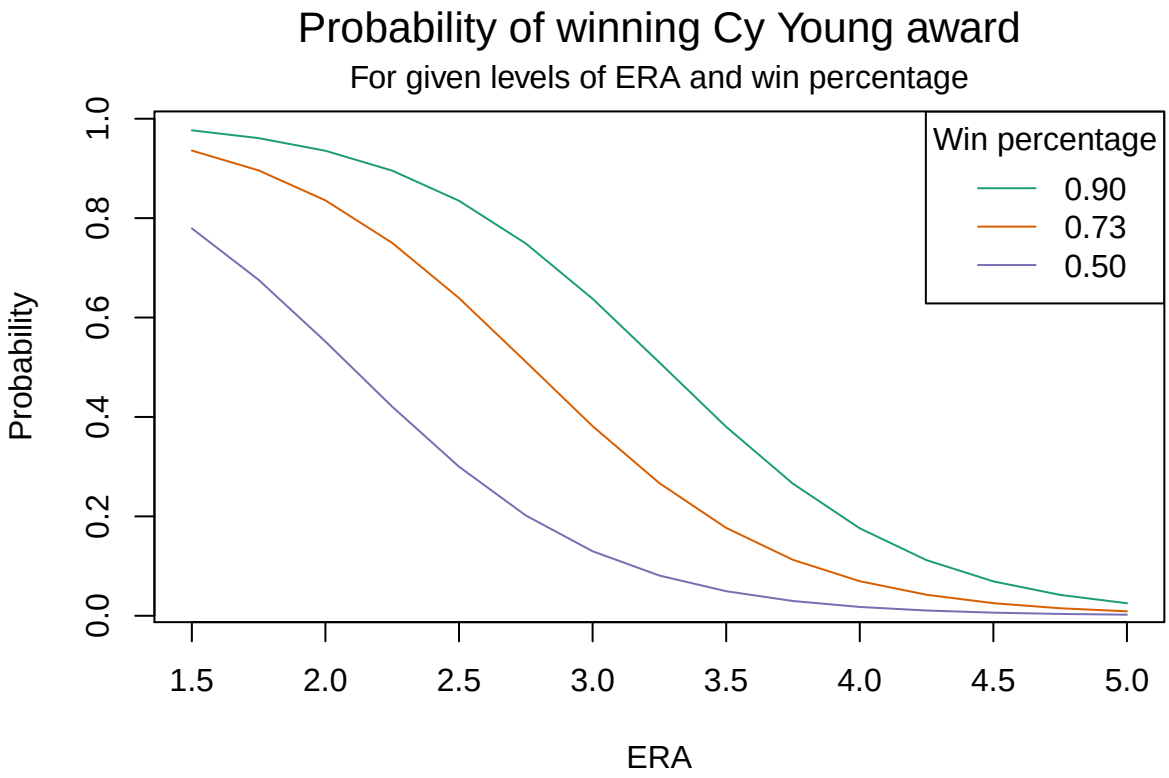
```
##              Estimate Std. Error
## (Intercept)  1.341659  3.2890963
## era         -2.109834  0.5126381
## winpct       6.170699  3.9103315
```

Question 1b

```
calc_probs <- function(era_val, winpct_val) {
  1 / (1 + exp(-1 * (coef(fit1)[1] + coef(fit1)[2]*era_val + coef(fit1)[3]*winpct_val)))
}

round(mean(df$winpct), digits = 2) # 0.73
```

```
## [1] 0.73
```

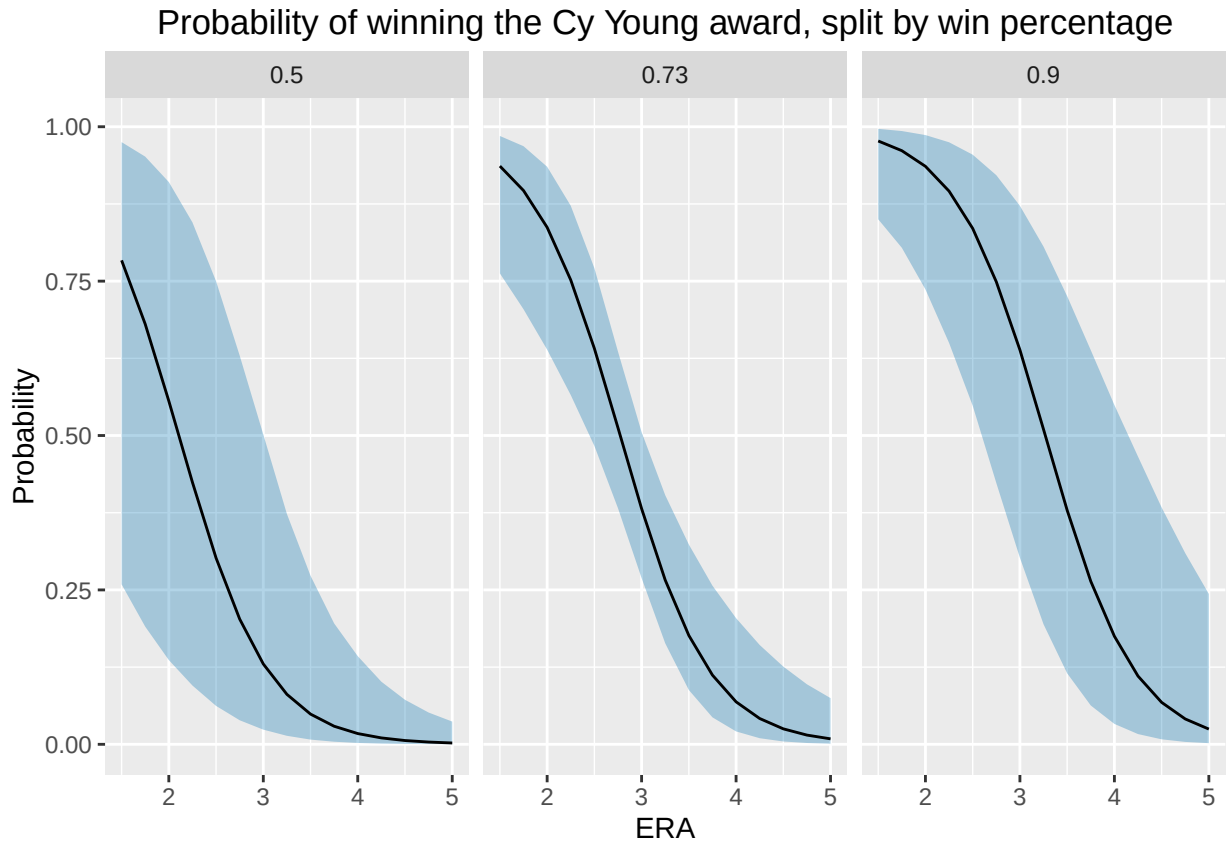
```
df2 <- data.frame(ERA = seq(1.5, 5, by = 0.25)) %>%
  mutate(
    pred_90 = mapply(calc_probs, ERA, 0.9),
    pred_mean = mapply(calc_probs, ERA, 0.73),
    pred_50 = mapply(calc_probs, ERA, 0.5)
  )

cols <- brewer.pal(3, "Dark2")

with(df2, plot(ERA, pred_90, type = "n", ylab = "Probability"))
mtext("Probability of winning Cy Young award", cex = 1.3, side = 3, line = 1.6)
mtext("For given levels of ERA and win percentage", side = 3, line = 0.4)

for (i in 1:3) lines(df2[, "ERA"], df2[, i+1], col = cols[i])

legend("topright", title = "Win percentage",
  legend = c("0.90", "0.73", "0.50"), lty = c(1,1,1), col = cols[1:3])
```



Question 1c

```
set.seed(456)

pe <- result1$par
vc <- solve(-1 * result1$hessian)

sims <- 10000
betas <- mvrnorm(sims, pe, vc)

predict_sims <- function(era, winpct) {
  pred <- sort(1 / (1 + exp(-1 * (betas[,1] + betas[,2]*era + betas[,3]*winpct))))
  quantile(pred, probs = c(0.5, 0.025, 0.975))
}

df3 <- expand.grid(era = seq(1.5, 5, by = 0.25), winpct = c(0.5, 0.73, 0.90))

df3[, c("pe", "ci_lo", "ci_hi")] <- t(mapply(
  FUN = predict_sims,
  era = df3$era, winpct = df3$winpct
))
```

```r
# plot the CIs
ggplot(data = df3, aes(x = era, y = pe)) +
  geom_ribbon(aes(ymin = ci_lo, ymax = ci_hi), alpha = 0.3, fill = "#0072B2") +
  geom_line() +
  facet_grid(. ~ winpct) +
  ggtitle("Probability of winning the Cy Young award, split by win percentage") +
  ylab("Probability") + xlab("ERA")
```



The figure above shows the probability of winning the Cy Young award across levels of ERA and the pitcher's personal win percentage. At a given level of win percentage, ERA is negatively associated with probability of winning the Cy Young award. At a given level of ERA, win percentage (at levels of 0.5, 0.73 and 0.9) is positively associated with probability of winning the Cy Young award. Of the three levels of win percentage shown, the confidence interval around the probability of winning the Cy Young award is tightest for the win percentage of 0.73.

Question 1d

I chose a model that additionally includes an indicator for whether the pitcher played in the National League. I expect this to improve the model because, due to a difference in rules, it is harder for American League pitchers to achieve low ERAs. This fact is probably taken into consideration in awarding the Cy Young, so it is important to include in the model.

I found mixed evidence that the model is improved with the indicator for whether the pitcher played in the American League. The likelihood ratio test was not statistically significant ($p = 0.15$). AIC was higher

(aic.test=0.025; bad), but BIC was also higher (bic.test=2.49; good). The AUC value was higher for the new model, both in-sample and after cross-validation. The actual-versus-predicted plots show no meaningful patterns.

```
set.seed(789)

library(nlme)
library(verification)
source("http://faculty.washington.edu/cadolph/software/avp.R")


y2 <- df$cy
x2 <- cbind(df$era, df$winpct, df$natleag)

stval2 <- lm(y2~x2)$coefficients

result2 <- optim(
  par = stval2, fn = llk.logit,
  method = "BFGS", hessian = T, control = list(fnscale = -1),
  y = y2, x = x2
)


pe.1 <- result1$par
vc.1 <- solve(-1 * result1$hessian)
se.1 <- sqrt(diag(vc.1))
ll.1 <- result1$value

pe.2 <- result2$par
vc.2 <- solve(-1 * result2$hessian)
se.2 <- sqrt(diag(vc.2))
ll.2 <- result2$value


# LR test
lr.test <- 2*(ll.2 - ll.1)
pchisq(lr.test,df=1,lower.tail=FALSE)
```

```
## [1] 0.1546861
```

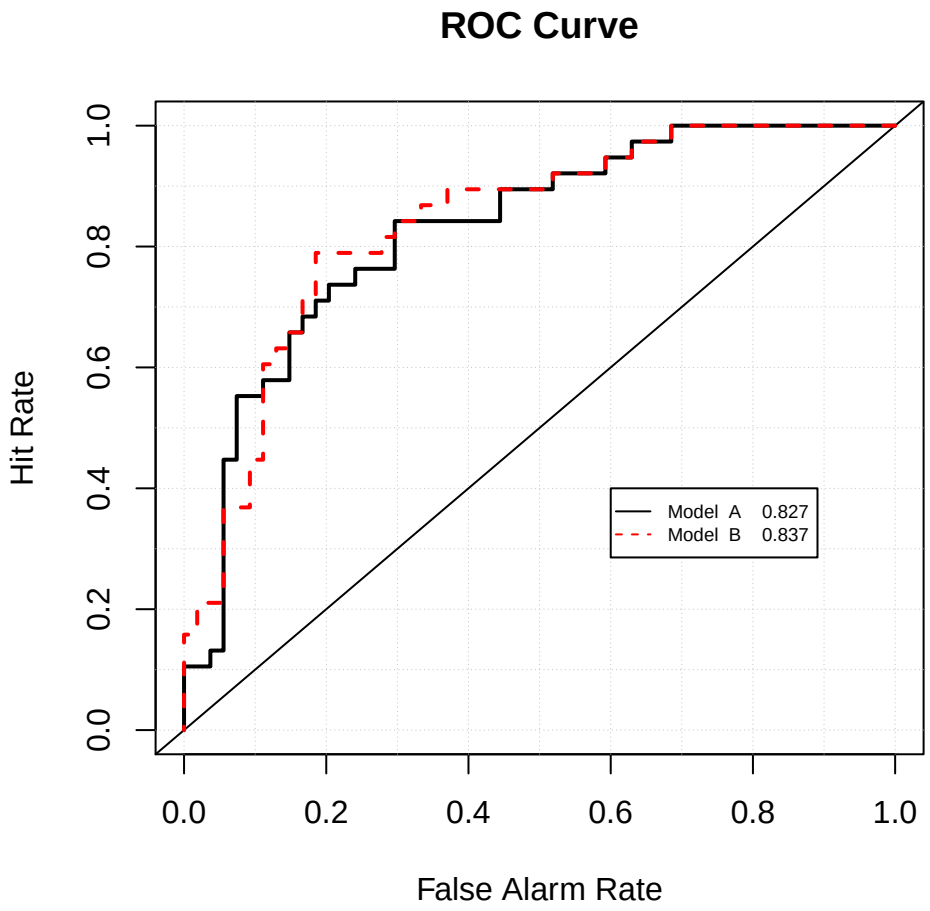
```
# AIC
2*(ll.2 - ll.1) - 1*2
```

```
## [1] 0.02542016
```

```
# BIC
-2*(ll.2 - ll.1) + 1*log(nrow(x))
```

```
## [1] 2.496368
```

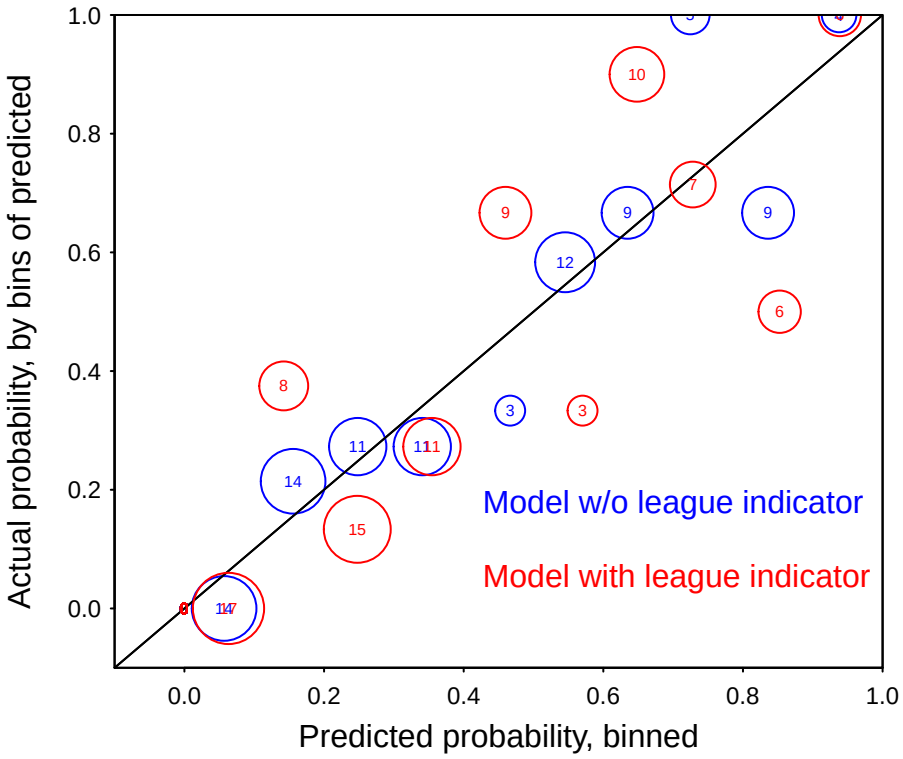
```r
# ROC
yhat.1 <- 1/(1+exp(-cbind(rep(1,nrow(x)),x)%*%pe.1))
yhat.2 <- 1/(1+exp(-cbind(rep(1,nrow(x2)),x2)%*%pe.2))
roc.plot(y,cbind(yhat.1, yhat.2), show.thres = FALSE, legend = TRUE)
```




```r
# actual v predicted plot

avp(y,
  x=cbind(rep(1,nrow(x)),x),
  beta=pe.1,
  fnform="logit",
  cutpoints=seq(0, 1, by = 0.1),
  usr=c(-0.1,1,-0.1,1),
  sizefactor=1.2,
  color = "blue",
  lab = list(x = .7, y=0.175, str="Model w/o league indicator", col="blue", cex=1),
  ylab = "Actual probability, by bins of predicted",
  xlab = "Predicted probability, binned" )
```

```
## [1] 0.7717391
```

```
avp(y2,
  x=cbind(rep(1,nrow(x2)),x2),
  beta=pe.2,
  fnform="logit",
  cutpoints=seq(0, 1, by = 0.1),
  usr=c(-0.1,1,-0.1,1),
  sizefactor=1.2,
  color = "red",
  lab = list(x = .705, y=0.05, str="Model with league indicator", col="red", cex=1),
  ylab = "Actual probability, by bins of predicted",
  xlab = "Predicted probability, binned",
  addtoplot = T)
```




```
## [1] 0.7608696
```

```
# cross-validation of AUC values

loocv <- function (obj) {
  data <- obj$data
  m <- dim(data)[1]
  form <- formula(obj)
  fam <- obj$family$family
```

```
  loo <- rep(NA, m)

  for (i in 1:m) {
    i.glm <- glm(form, data = data[-i, ], family = fam)
    loo[i] <- predict(i.glm, newdata = data[i,], family = fam, type = "response")
  }

  loo
}

model1 <- glm(cy ~ era + winpct, data = df, family = binomial)
model2 <- glm(cy ~ era + winpct + natleag, data = df, family = binomial)

yhat.1.cv <- loocv(model1)
yhat.2.cv <- loocv(model2)


cat(paste(
  "Model 1 AUC:", round(roc.area(y, yhat.1)$A, digits = 4),
  "\nModel 2 AUC:", round(roc.area(y2, yhat.2)$A, digits = 4)
))
```

```
## Model 1 AUC: 0.8275
## Model 2 AUC: 0.8372
```

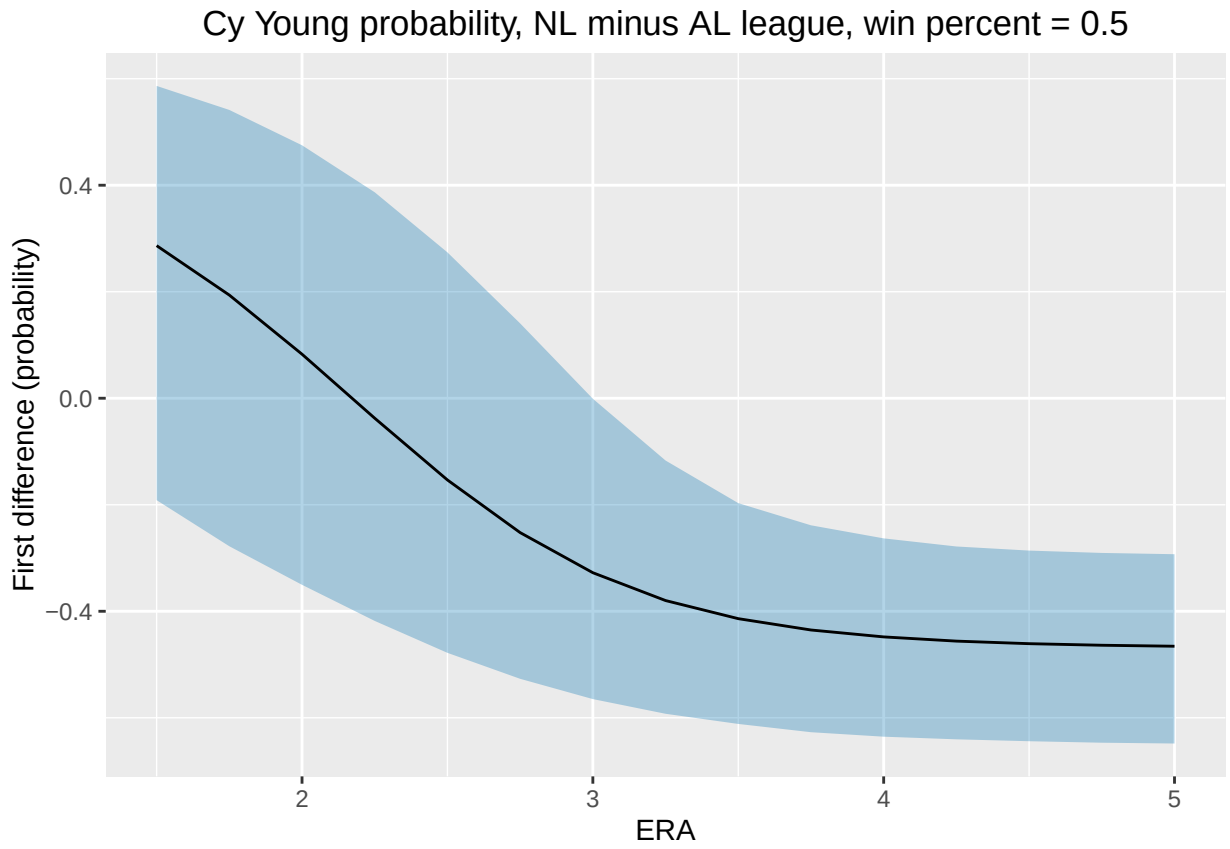
```
cat(paste(
  "Model 1 AUC:", round(roc.area(y, yhat.1.cv)$A, digits = 4),
  "\nModel 2 AUC:", round(roc.area(y2, yhat.2.cv)$A, digits = 4)
))
```

```
## Model 1 AUC: 0.7987
## Model 2 AUC: 0.8036
```

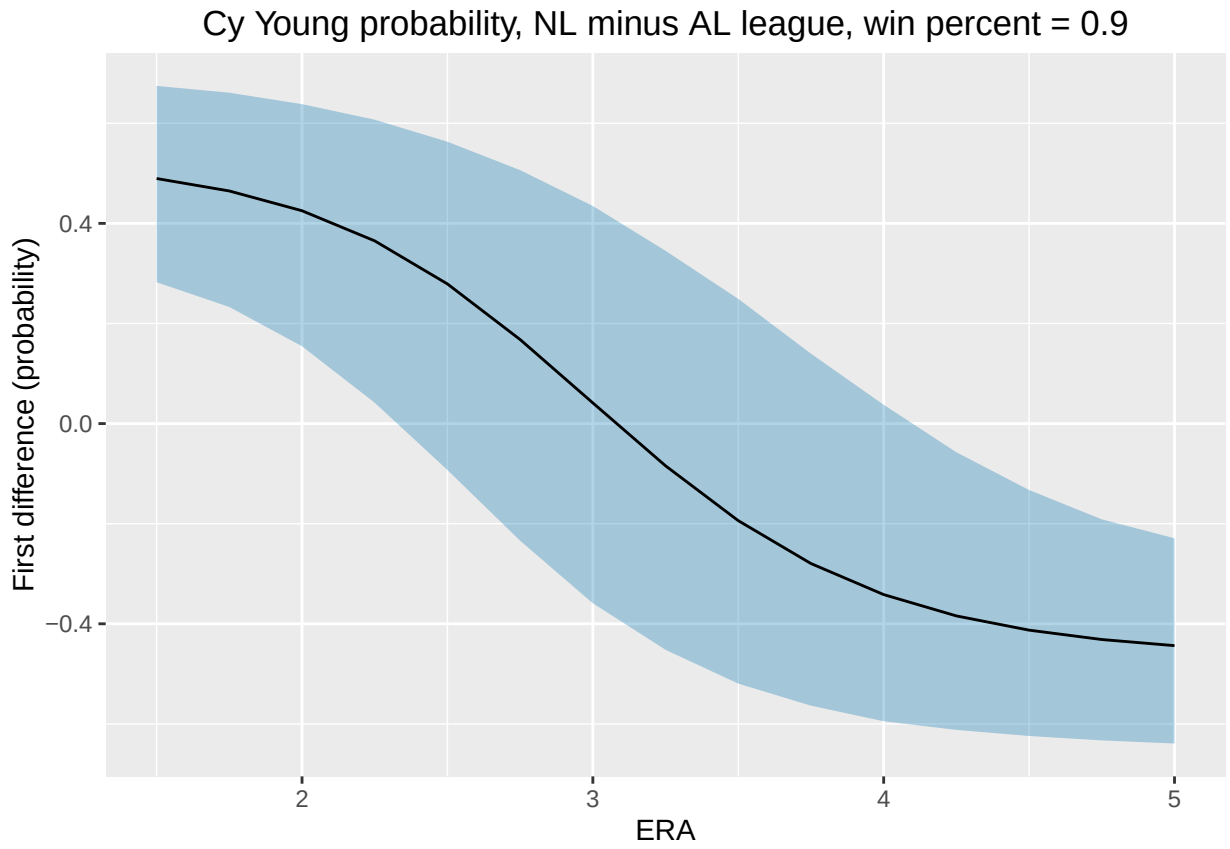
Question 1e

```
y2 <- df$cy
x2 <- cbind(df$era, df$winpct, df$natleag)

stval2 <- lm(y2~x2)$coefficients

result2 <- optim(
  par = stval2, fn = llk.logit,
  method = "BFGS", hessian = T, control = list(fnscale = -1),
  y = y2, x = x2
)

pe.2 <- result2$par
vc.2 <- solve(-1 * result2$hessian)
```

```r
se.2 <- sqrt(diag(vc.2))
ll.2 <- result2$value

# scenario 1, first difference for leagues (NL, AL) across levels of ERA
#   with winpct set to 0.5
xscen <- cfMake(
  formula = cy ~ era + winpct + natleag,
  data = df,
  nscen = 15
)

era_vec <- seq(1.5, 5, by = 0.25)

for (i in 1:length(era_vec)) {
  xscen <- cfChange(xscen, "era", x = era_vec[i], scen = i)
  xscen <- cfChange(xscen, "winpct", x = 0.5, scen = i)
  xscen <- cfChange(xscen, "natleag", x = 1, xpre = 0, scen = i)
}

sims <- 10000
betas2 <- mvrnorm(sims, pe.2, vc.2)

df4 <- as.data.frame(logitsimfd(xscen, betas2, ci=0.95)) %>%
  mutate(era = era_vec)

ggplot(data = df4, aes(x = era, y = pe)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.3, fill = "#0072B2") +
  geom_line() +
  ggtitle("Cy Young probability, NL minus AL league, win percent = 0.5") +
  ylab("First difference (probability)") + xlab("ERA")
```




```
# scenario 2, first difference for win percent (0.83, 0.73) across levels of ERA
#   with winpct set to 0.9
xscen2 <- cfMake(
  formula = cy ~ era + winpct + natleag,
  data = df,
  nscen = 15
)

for (i in 1:length(era_vec)) {
  xscen2 <- cfChange(xscen2, "era", x = era_vec[i], scen = i)
  xscen2 <- cfChange(xscen2, "winpct", x = 0.9, scen = i)
  xscen2 <- cfChange(xscen2, "natleag", x = 1, xpre = 0, scen = i)
}

df5 <- as.data.frame(logitsimfd(xscen2, betas2, ci=0.95)) %>%
  mutate(era = era_vec)

ggplot(data = df5, aes(x = era, y = pe)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.3, fill = "#0072B2") +
  geom_line() +
  ggtitle("Cy Young probability, NL minus AL league, win percent = 0.9") +
  ylab("First difference (probability)") + xlab("ERA")
```



Question 1f

Logistic regression does offer a defensible probability model for this analysis. It bounds predictions between 0 and 1, so the predicted probabilities will always be interpretable. Also, logit model assumes that the probability distribution is symmetrical. The sample appears roughly evenly split between Cy Young winners and losers, so symmetry is an acceptable assumption. One potential problem is that "cy" observations are not independent, because some pitchers appear multiple times in the dataset. This artificially decreases standard errors leading to anti-conservative hypothesis testing.

Question 1g

Excluding pitchers unlikely to win the award would change the regression coefficients, but theoretically (assuming that the functional form of the model is correct) it shouldn't change the predicted probabilities. A pitcher with ERA=2 and winpct=0.8 should have the same predicted probability of winning the Cy Young with both data sets. One difference is that there would be no in-sample predictions for worse levels of ERA and winpct. Also, if the new data set is a subset of the old one, standard errors would increase and may affect conclusions from tests of statistical significance.