PS3 Solutions

Teaching Staff

February 2023

Q1

```
set.seed(123)
x = rexp(1500, rate = 2)
```

1

```
boot_univariate = function(datvec, statint, B, alpha){
  out = vector(mode = "logical", length = B)
  for(i in 1:B){
    out[i] = statint(sample(datvec, replace = T))
  }
  conf.out = quantile(out, probs = c(alpha/2, 1-alpha/2))
  return(conf.out)
}
```

2

```
boot_univariate(x, median, 10000, 0.05)
```

```
##      2.5%     97.5% 
## 0.3313953 0.3856648
```

Bonus

```
iqr = function(x){
  return(quantile(x, probs = .75) - quantile(x, probs = .25))
}
boot_univariate(x, iqr, 10000, 0.05)
```

```
##      2.5%     97.5% 
## 0.5170515 0.6014724
```

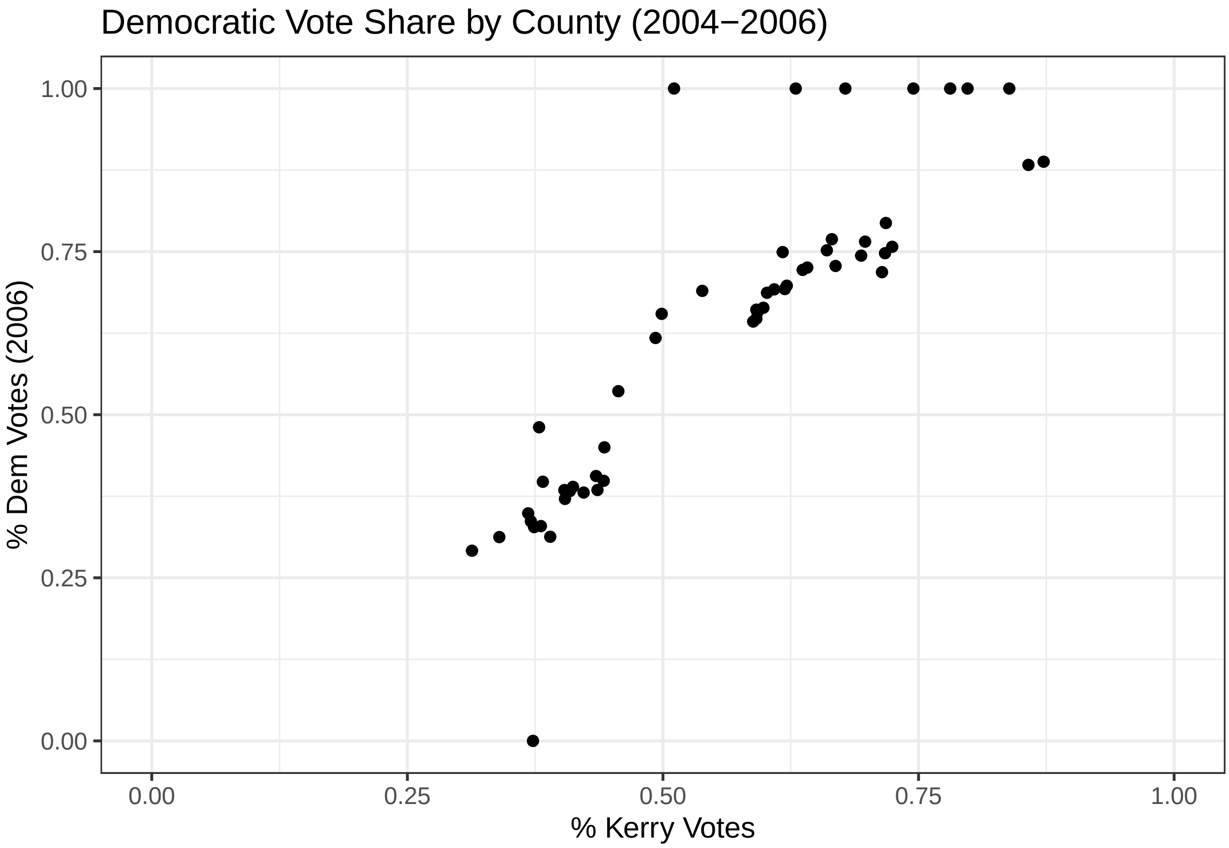
Q2

1

```
ca = read.csv("ca2006.csv")
```

2

```
plot = ca |>
  ggplot(aes(dem_pres_2004, prop_d))+
  geom_point()+
  labs(x = "% Kerry Votes",
       y = "% Dem Votes (2006)",
       title = "Democratic Vote Share by County (2004-2006)")+
  theme_bw()+
  ylim(0,1)+
  xlim(0,1)
plot
```



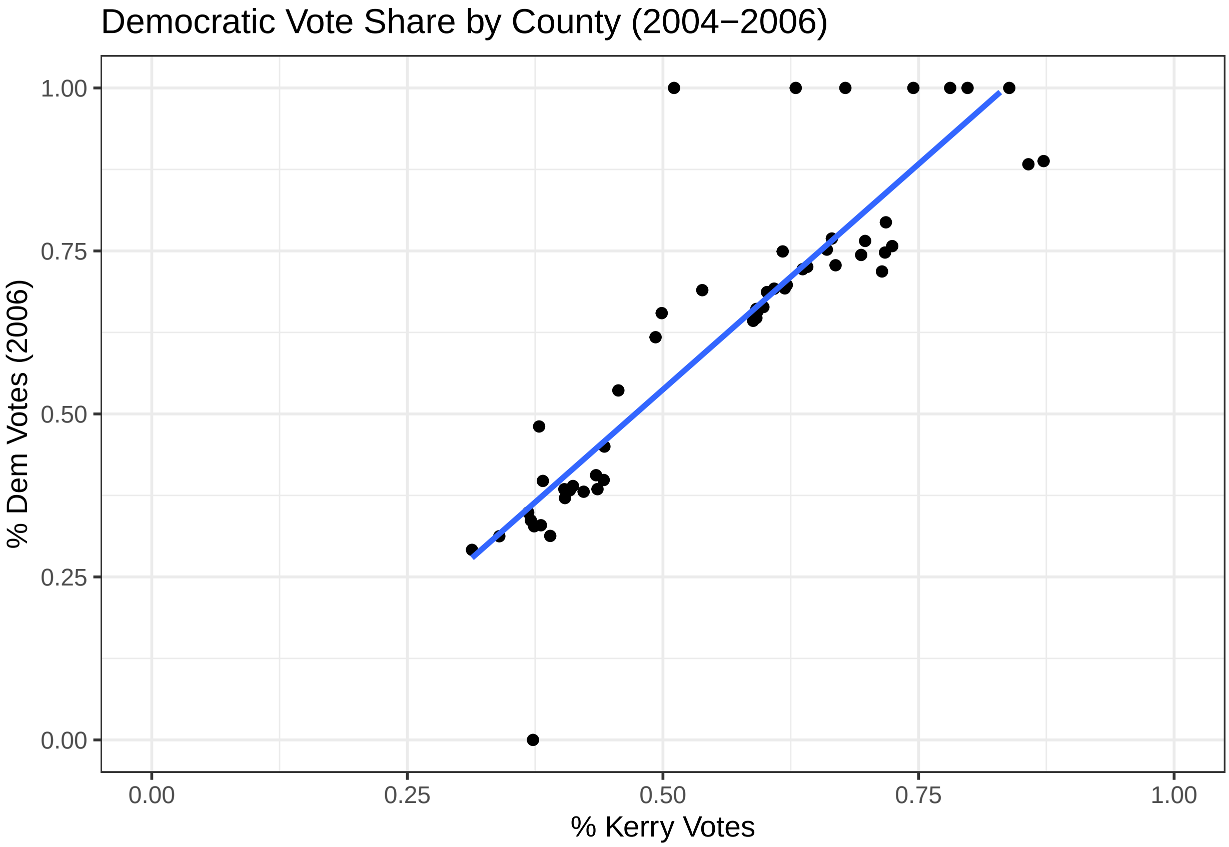
3

```
reg = lm(prop_d ~ dem_pres_2004, data = ca)
summary(reg)
```

```
##
## Call:
## lm(formula = prop_d ~ dem_pres_2004, data = ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36168 -0.04314 -0.00830  0.01233  0.44754
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.15390    0.05978  -2.574    0.013 *
## dem_pres_2004  1.38268    0.10291  13.436   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1125 on 51 degrees of freedom
## Multiple R-squared:  0.7797, Adjusted R-squared:  0.7754
## F-statistic: 180.5 on 1 and 51 DF,  p-value: < 2.2e-16
```

```
plot + geom_smooth(method = "lm", se = F)
```



4

```
my_predict = function(coefs, newdata, ols = TRUE){
  if(ols == TRUE){
    ## Linear Model prediction
    prediction = sum(coefs%*%newdata)
    return(unname(prediction))
  }else{
    ## Logit Model prediction
    betas = unname(coefs) %*% newdata
    odds = 1/ (1 + exp(-betas))
    return(odds)
  }
}
```

```r
my_predict(reg$coefficients, newdata = c(1, 0.5))
```

```
## [1] 0.5374445
```

5 and 6

```r
mreg = lm(prop_d ~ dem_pres_2004 + dem_pres_2000 + dem_inc,
          data = ca)
my_predict(mreg$coefficients, newdata = c(1,0.5, 0.5, 1), ols = TRUE)
```

```
## [1] 0.6147444
```

7

```r
boot_reg = function(df, N = 53, B = 10000, alpha = 0.05){
  set.seed(pi)
  simple = vector(mode = "logical", length = B)
  multi = vector(mode = "logical", length = B)
  for(i in 1:B){
    dat = df[sample.int(nrow(df), 53, replace = T),]
    simple[i] = my_predict(lm(prop_d ~ dem_pres_2004,
                              data = dat)$coefficients,
                           newdata = c(1,0.5))
    multi[i] = my_predict(lm(prop_d ~ dem_pres_2004 +
                               dem_pres_2000 + dem_inc,
                             data = dat)$coefficients,
                          newdata = c(1,0.5,0.5,1))
  }
  sci = quantile(simple, probs = c(alpha/2, 1-alpha/2))
  mci = quantile(multi, probs = c(alpha/2, 1-alpha/2))
  return(list(simple = simple, multi = multi,
              simple_ci = sci, multi_ci = mci))
}
```

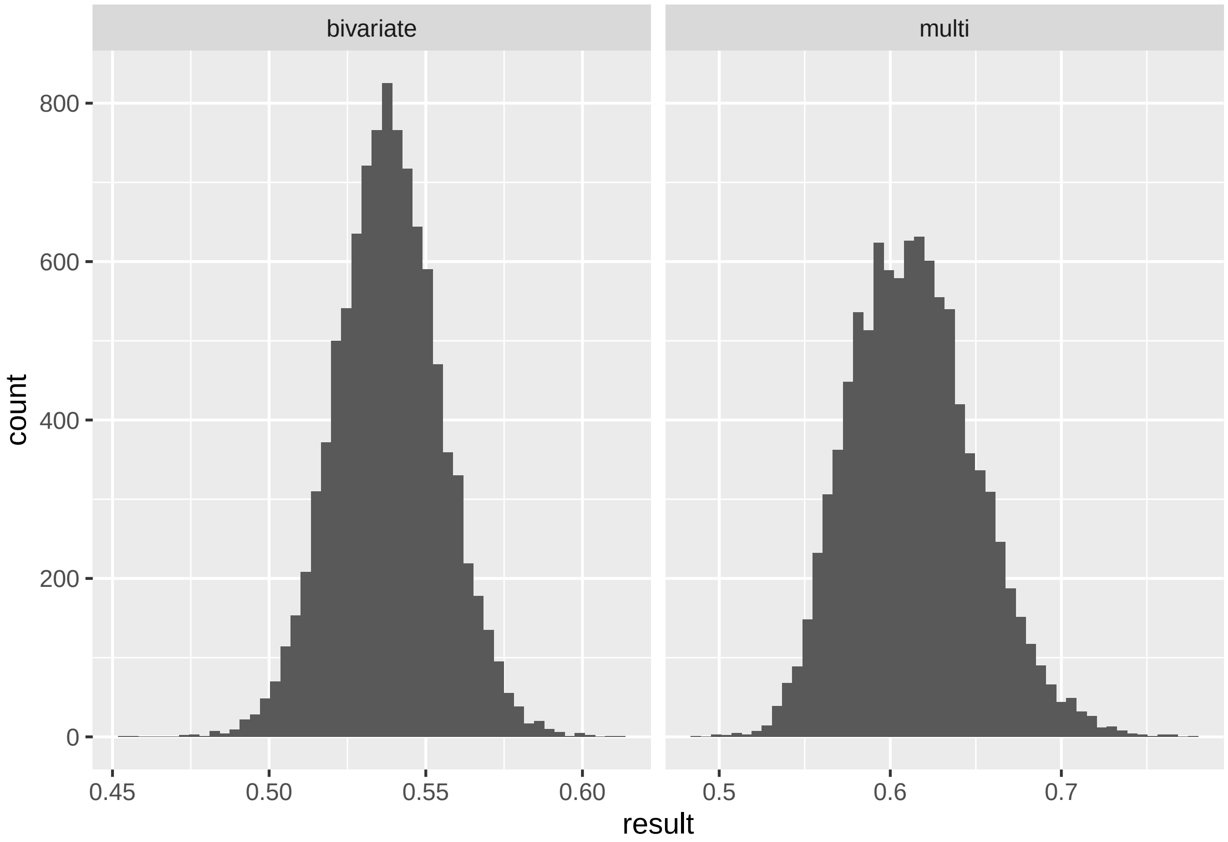
8

```r
results = boot_reg(df = ca)
```

2.5%	97.5%
0.5050168	0.5716776
0.5496060	0.6924033

```r
out = data.frame(id = c(rep("bivariate", 10000),
                          rep("multi", 10000)),
                   result = c(results$simple,
                              results$multi))
```

```r
out |>
  ggplot(aes(result))+
  geom_histogram(bins = 50)+
  facet_wrap(~id, scales = "free_x")
```



9

```
mean(results$simple > .5)
```

```
## [1] 0.988
```

```
mean(results$multi > .5)
```

```
## [1] 0.9996
```

Q3

1 and 2

```
clinton = read.csv("vote92.csv")
mean(clinton$clintonvote)
```

```
## [1] 0.4576458
```

3

```
logit = glm(clintonvote ~ dem + female + clintondist, data = clinton,
            family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = clintonvote ~ dem + female + clintondist, family = "binomial",
##     data = clinton)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9866  -0.6183  -0.3559   0.6317   2.7461
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.40692    0.18758  -7.500 6.37e-14 ***
## dem          3.05648    0.18687  16.357  < 2e-16 ***
## female       0.17417    0.18413   0.946    0.344
## clintondist -0.14482    0.02777  -5.215 1.84e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1253.61  on 908  degrees of freedom
## Residual deviance:  769.17  on 905  degrees of freedom
## AIC: 777.17
##
## Number of Fisher Scoring iterations: 5
```

4 and 5

```r
## see my_predict() function definition
my_predict(logit$coefficients, newdata = c(1,1,1,1), ols = FALSE)
```

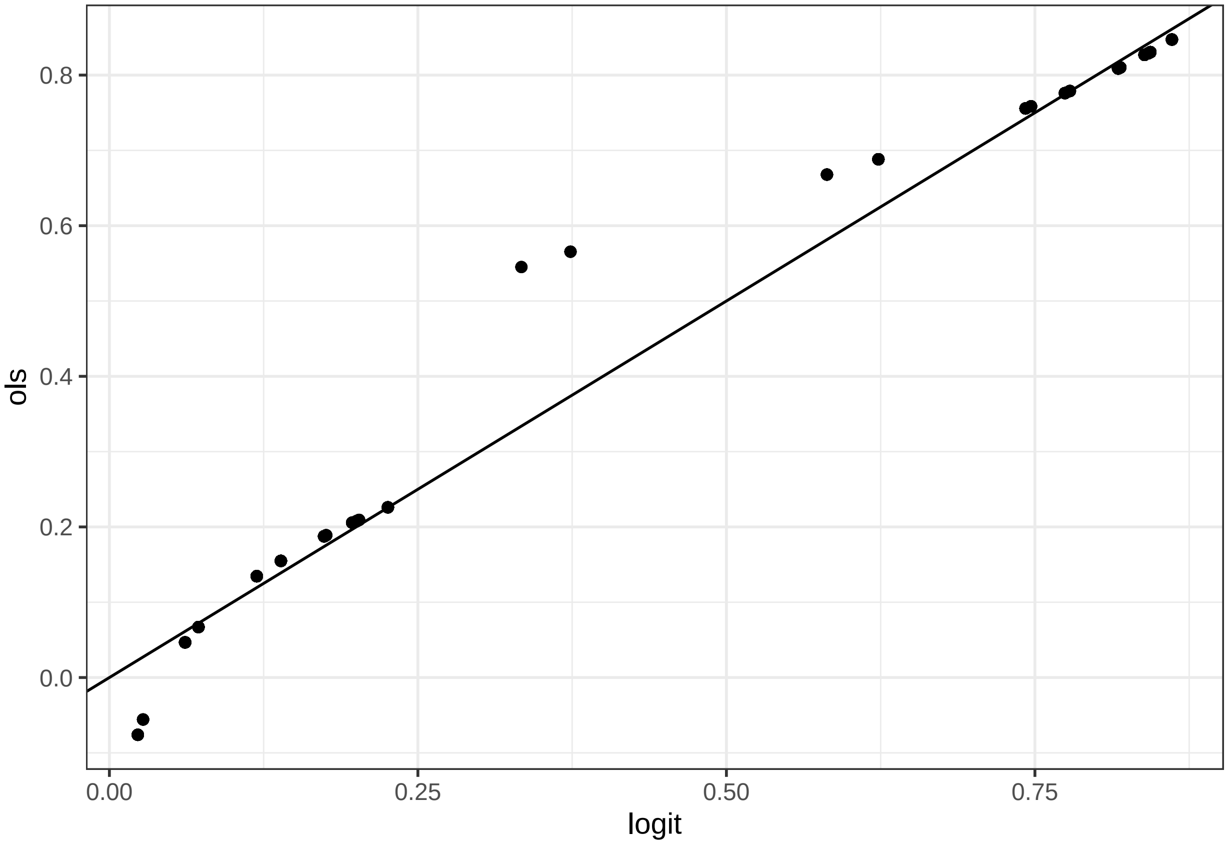
```
##           [,1]
## [1,] 0.8427606
```

6

```r
ols = lm(clintonvote ~ dem + female + clintondist, data = clinton)
ols.preds = vector(mode = "logical", nrow(clinton))
logit.preds = vector(mode = "logical", nrow(clinton))

for(i in 1:nrow(clinton)){
  newdata = newdata = c(1, as.numeric(clinton[i,c(2:4)]))
  ols.preds[i] = my_predict(ols$coefficients, newdata, ols = TRUE)
  logit.preds[i] = my_predict(logit$coefficients, newdata, ols = FALSE)
}
```

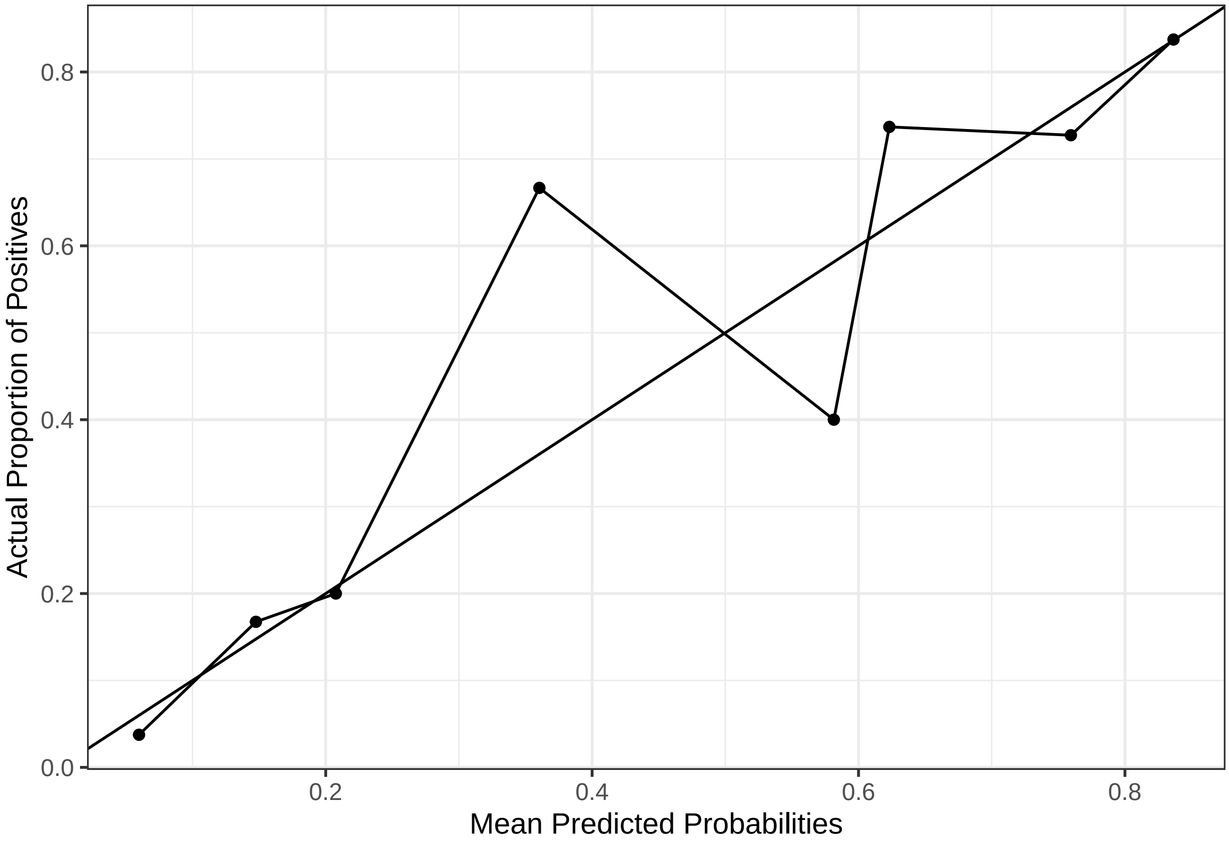
```r
data.frame(ols = ols.preds, logit = logit.preds) |>
  ggplot(aes(logit, ols))+
  geom_point()+
  geom_abline(intercept = 0,slope = 1)+
  theme_bw()
```



Bonus

```
bins = cut(logit.preds, breaks = seq(0,1,.1), right = FALSE,
           labels = c(1:10))
bonusDat = data.frame(preds = logit.preds, bins = bins)
mean_prob = aggregate(bonusDat$preds, by = list(bins), FUN=mean)
posi_prob = aggregate(clinton$clintonvote, by = list(bins), FUN=mean)

data.frame(mean_prob = mean_prob$x, posi_prob = posi_prob$x) |>
  ggplot(aes(mean_prob, posi_prob))+
  geom_point()+
  geom_line()+
  geom_abline(intercept = 0, slope = 1)+
  theme_bw()+
  labs(x = "Mean Predicted Probabilities",
       y = "Actual Proportion of Positives")
```


Actual Proportion of Positives
0.0
0.2
0.4
0.6
0.8
0.2
0.4
0.6
0.8
Mean Predicted Probabilities