Problem Set 8

Daniel Shapiro

11/6/2022

Question 1 Background:

Download the x.csv dataset from the course website. This data contains a set of fixed values for an independent variable X. Consider the following population regression model, where u is the error term:

$$\begin{aligned} y_i &= 3 + 5x_i + u_i \\ u_i &\sim N(0,1) \end{aligned}$$

In this situation we know the true population parameters $\beta_0 = 3$ and $\beta_1 = 5$.

```
data <- read.csv("x.csv")
```

1a) Simulate the sampling distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ by doing the following steps $m = 1000$ times:

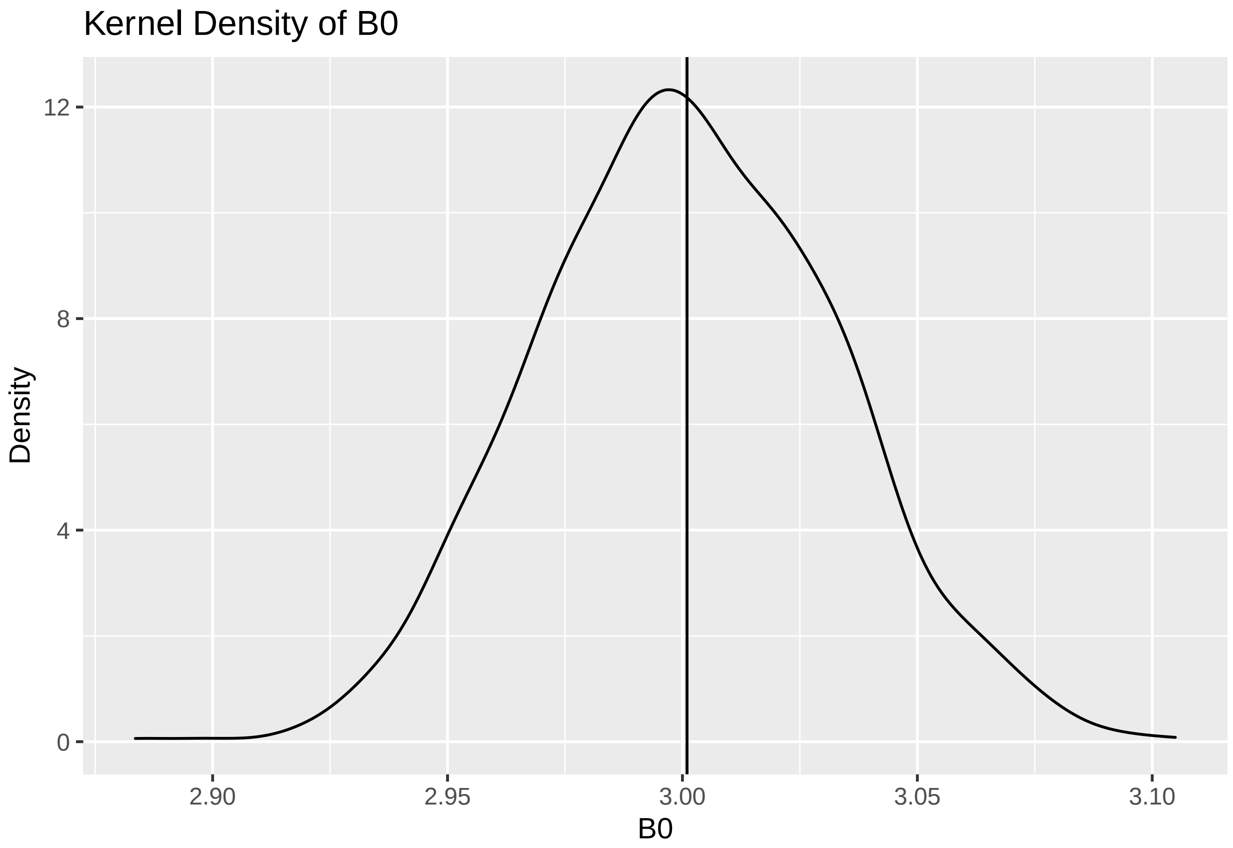
1. Generate random errors u from the N(0,1) distribution.
2. Generate values for y using u, the fixed x, and the true population parameters.
3. Run a regression of y on x.
4. Record your OLS estimates.
5. Repeat.

At the end of this process, you should have m draws of $\hat{\beta}_0$ and $\hat{\beta}_1$ which serves as draws from your sampling distributions. Generate a kernel density plot for your two sampling distributions. Superimpose a line on each for the mean of the distributions. From your simulations, does the OLS estimator appear to be unbiased? Do the standard errors you get from the individual regressions match up to what you find from the sampling distributions?

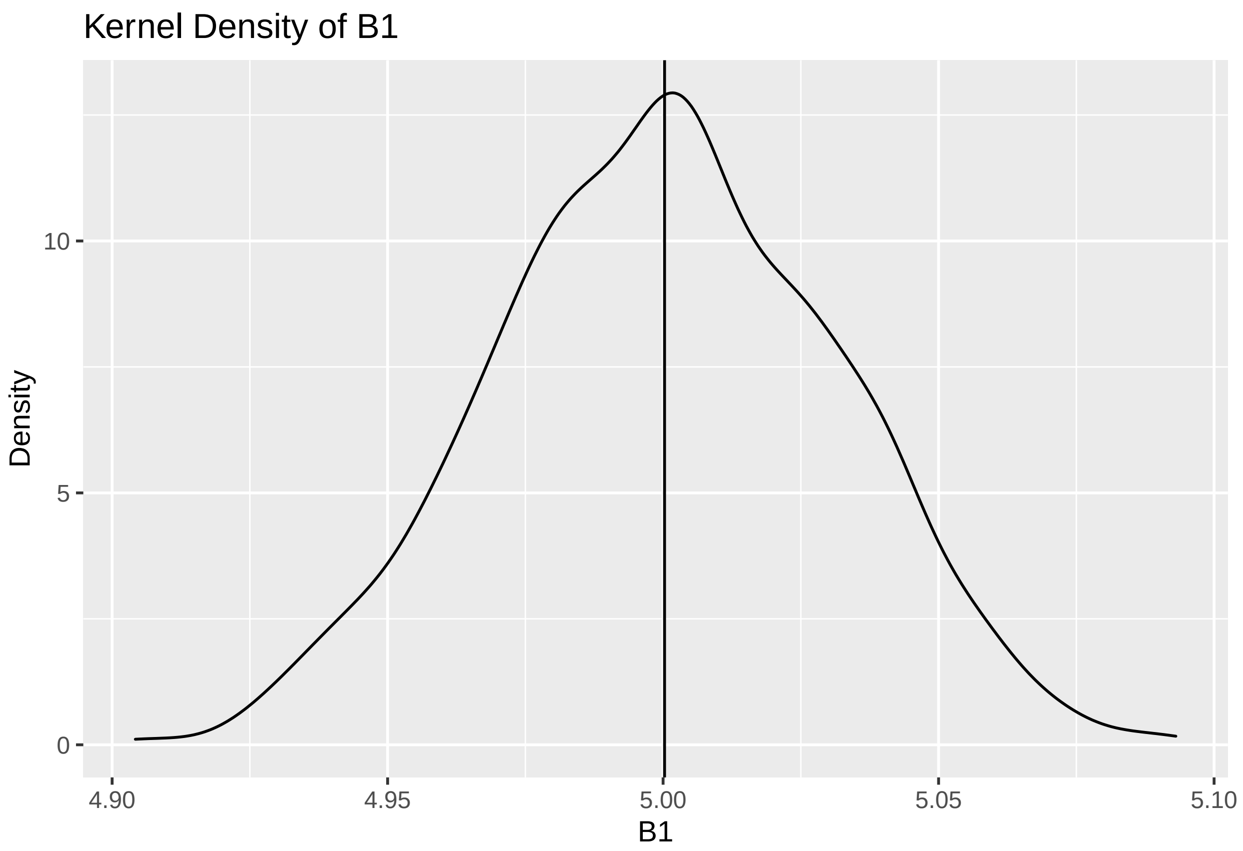
```
dataframe <- data.frame(matrix(ncol = 3, nrow = 1000))

for(i in 1:1000){
maindata <- data %>%
  mutate(u = rnorm(1000)) %>%
  mutate(y = (3 + 5*`x` + `u`))
```

```r
regression <- lm(y ~ x, data = maindata)

m <- summary(regression)

dataframe$X1[i] <- regression$coefficients[1]

dataframe$X2[i] <- regression$coefficients[2]

dataframe$X3[i] <- m$sigma
}
```

```r
ggplot(dataframe, aes(X1)) +
  geom_density() +
  labs(x = "B0",
       y = "Density",
       title = "Kernel Density of B0") +
  geom_vline(xintercept = mean(dataframe$X1))
```



```r
ggplot(dataframe, aes(X2)) +
  geom_density() +
  labs(x = "B1",
       y = "Density",
       title = "Kernel Density of B1") +
  geom_vline(xintercept = mean(dataframe$X2))
```

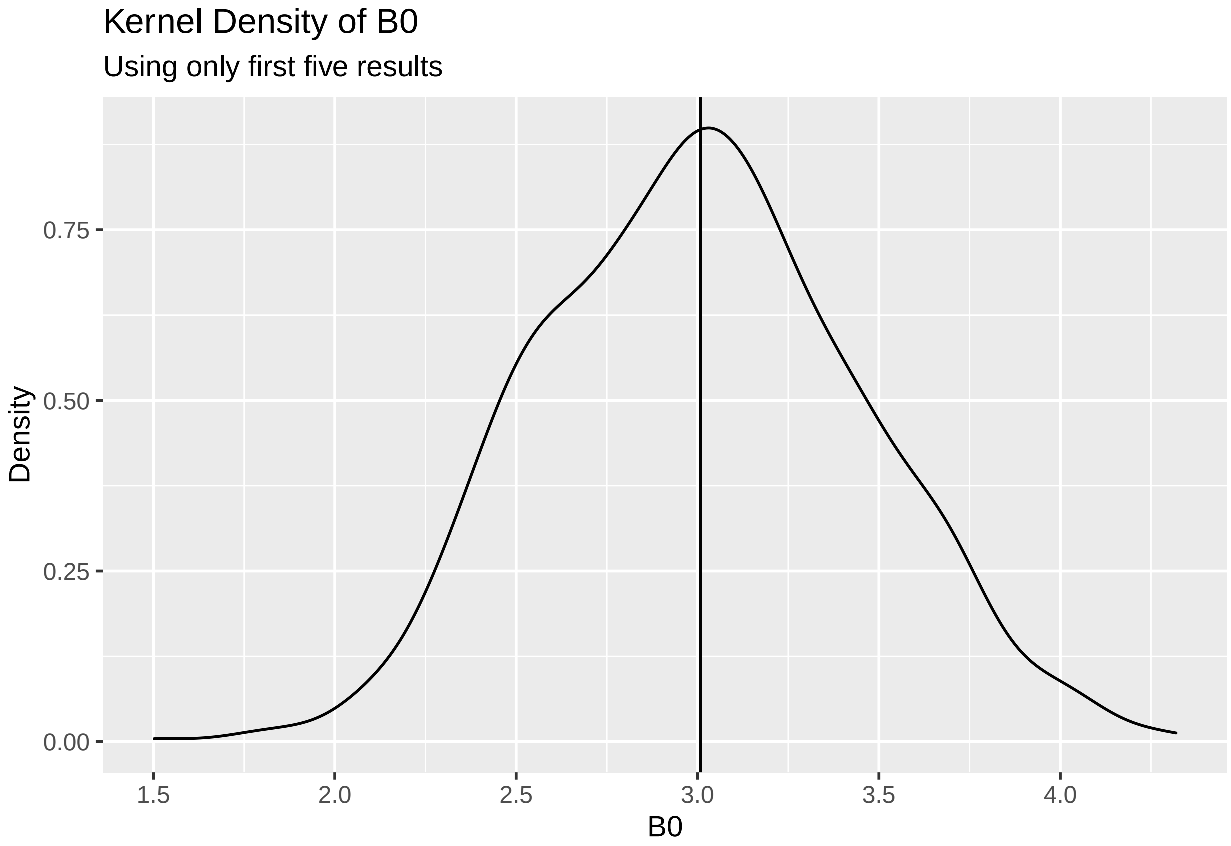


The density plots look fairly evenly spread/normally distributed around the sample means, which appear quite close to the original "3" and "5" coefficients (β_0 and β_1). So they do appear to be relatively unbiased. I also added a column "X3" to the dataframe (see my for loop) that pulls the standard error for each individual regression. They look to be all around 1, for the most part, which matches up.

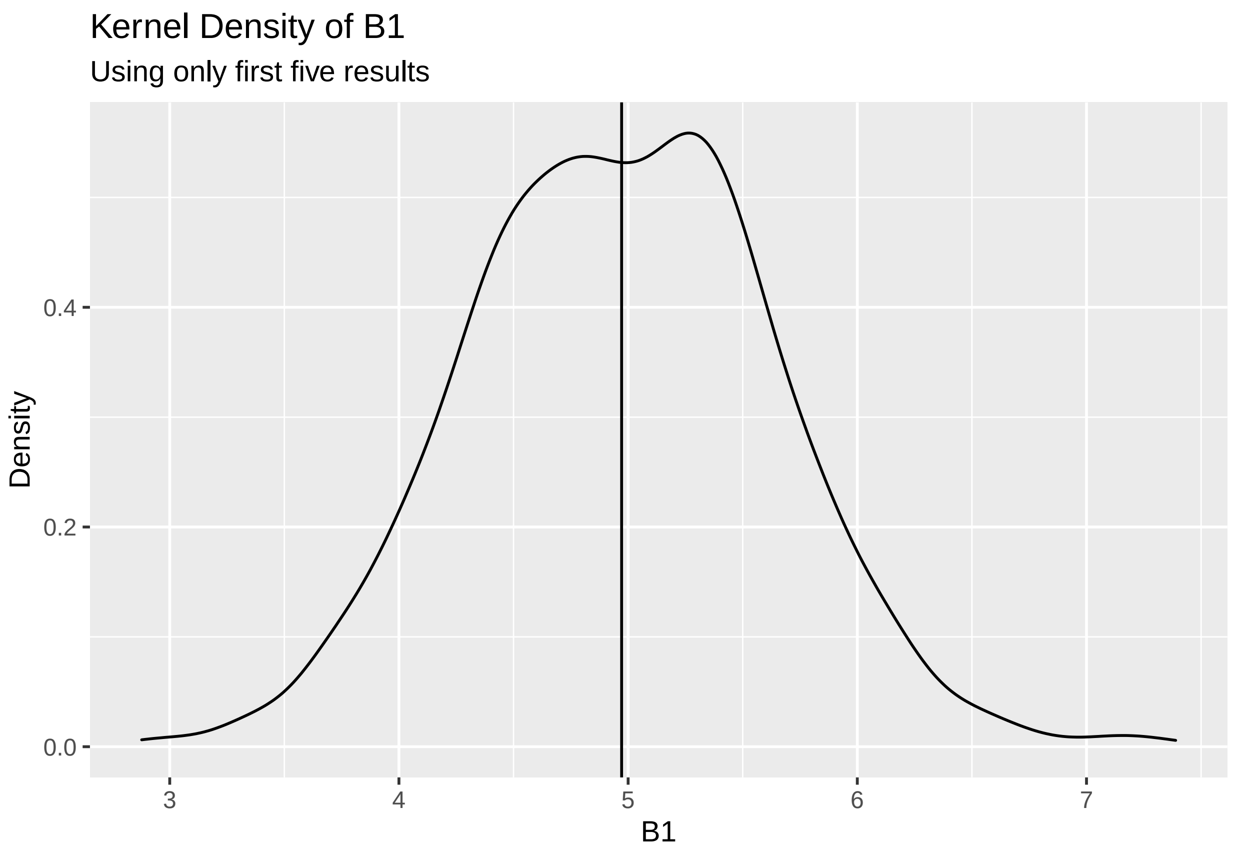
1b) Repeat (a), this time using just the first five observations of x ($n = 5$). How do your results compare? Why?

```
dataframe2 <- data.frame(matrix(ncol = 3, nrow = 1000))

for(i in 1:1000){
maindata <- data %>%
  mutate(u = rnorm(1000)) %>%
  mutate(y = (3 + 5*`x` + `u`))

bdata <- maindata[1:5,]

regression <- lm(y ~ x, data = bdata)

m <- summary(regression)

dataframe2$X1[i] <- regression$coefficients[1]

dataframe2$X2[i] <- regression$coefficients[2]
dataframe2$X3[i] <- m$sigma
}
```

```
ggplot(dataframe2, aes(X1)) +
  geom_density() +
  labs(x = "B0",
       y = "Density",
       title = "Kernel Density of B0",
       subtitle = "Using only first five results") +
  geom_vline(xintercept = mean(dataframe2$X1))
```



```
ggplot(dataframe2, aes(X2)) +
  geom_density() +
  labs(x = "B1",
       y = "Density",
       title = "Kernel Density of B1",
       subtitle = "Using only first five results") +
  geom_vline(xintercept = mean(dataframe2$X2))
```



There are certainly differences here. Both values have a much wider distribution; the X-axis is much more expansive than in the one in which we used all 1000 observations. This obviously makes sense because our n is smaller so there is more variance. The X3 column (standard error) still appears to be centered somewhere around 1 (I checked, and it's about .91), but there's a ton more variance. This all fits with the idea that we're using a smaller sample size so there's more room for variation.

1c) Repeat (a) and (b) except in this case, generate u from a uniform distribution ranging from -1 to 1. How does this change your results? Why?

```
dataframe3 <- data.frame(matrix(ncol = 3, nrow = 1000))

for(i in 1:1000){
maindata <- data %>%
  mutate(u = runif(n = 1000, min = -1, max = 1)) %>%
  mutate(y = (3 + 5*`x` + `u`))

regression <- lm(y ~ x, data = maindata)

m <- summary(regression)

dataframe3$X1[i] <- regression$coefficients[1]

dataframe3$X2[i] <- regression$coefficients[2]
```

```r
dataframe3$X3[i] <- m$sigma
}

dataframe4 <- data.frame(matrix(ncol = 3, nrow = 1000))

for(i in 1:1000){
maindata <- data %>%
  mutate(u = runif(n = 1000, min = -1, max = 1)) %>%
  mutate(y = (3 + 5*`x` + `u`))

ddata <- maindata[1:5,]

regression <- lm(y ~ x, data = ddata)

m <- summary(regression)

dataframe4$X1[i] <- regression$coefficients[1]

dataframe4$X2[i] <- regression$coefficients[2]

dataframe4$X3[i] <- m$sigma
}
```

The results for β_0 and β_1 don't really change here; the graphs look like they did in 1a) and 1b) – x-axis ranges and all. This is because we're only changing u_i, and β_0 and β_1 don't really depend on the error term. What does change is the third column that I created – the standard error. The standard error of each regression does have to do with the value of u in the setup, and a uniform distribution looks very different and provides very different values than the normal distribution. Thus it's no surprise that the X3 columns are centered closer to a different value (somewhere around .56 or so).

Question 2 Background:

In 1977, Douglas Hibbs published a paper called "Political Parties and Macroeconomic Policy" in which he analyzed the connections between the ideological orientation of governments and the results of their economic policy. You can find the data he used in hibbs.csv. He coded the percentage of years (out of 1945-69 period) Leftist parties had been in power (or had shared power as members of coalition governments) in 12 Western European and North American countries (percleft). He also coded the average inflation and average unemployment in these countries over the same interval. He was interested in the "revealed preference" of leftist governments to please their constituents with high-inflation, low-unemployment economic policy and vice versa for rightist governments. We will replicate his analysis here.

```r
hibbs <- read.csv("hibbs.csv")
```

We will run two separate bivariate regressions. In each regression, interpret what the slope and the intercept mean for the relationship between political parties and economic policy and also interpret the R^2. Interpret the statistical and practical significance, too. Plot a scatterplot of each with the regression line fitted onto the plot. Discuss the plausibility of each of the regression assumptions. Do you think each of the assumptions is valid?

2a) Run a regression of unemployment on the independent variable government ideology.

```
reg2a <- lm(unemployment ~ percleft, data = hibbs)
summary(reg2a)
```

```
##
## Call:
## lm(formula = unemployment ~ percleft, data = hibbs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5907 -0.5463  0.1795  0.8165  1.1758
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3.92421    0.59881   6.553 0.0000645 ***
## percleft    -0.03017    0.01022  -2.952    0.0145 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.115 on 10 degrees of freedom
## Multiple R-squared:  0.4657, Adjusted R-squared:  0.4123
## F-statistic: 8.717 on 1 and 10 DF,  p-value: 0.01447
```

2b) Run a regression of inflation on the independent variable government ideology.

```
reg2b <- lm(inflation ~ percleft, data = hibbs)
summary(reg2b)
```

```
##
## Call:
## lm(formula = inflation ~ percleft, data = hibbs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92376 -0.41876 -0.09741  0.52854  1.22631
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 2.632811   0.359722   7.319 0.0000254 ***
## percleft    0.020584   0.006138   3.353   0.00733 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6698 on 10 degrees of freedom
## Multiple R-squared:  0.5293, Adjusted R-squared:  0.4822
## F-statistic: 11.24 on 1 and 10 DF,  p-value: 0.007325
```