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")
```

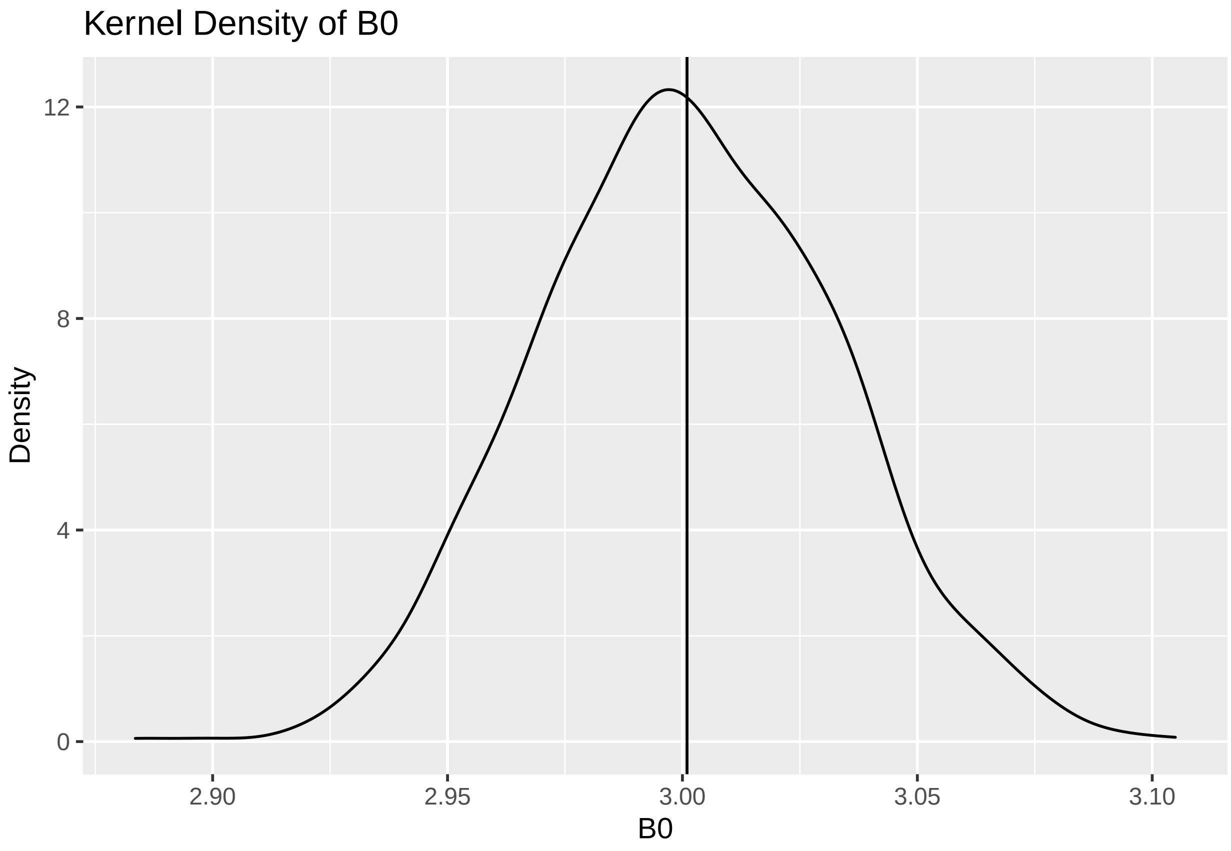
1a) Simulate the sampling distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ by doing the following steps $m = 1000$ times:

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.

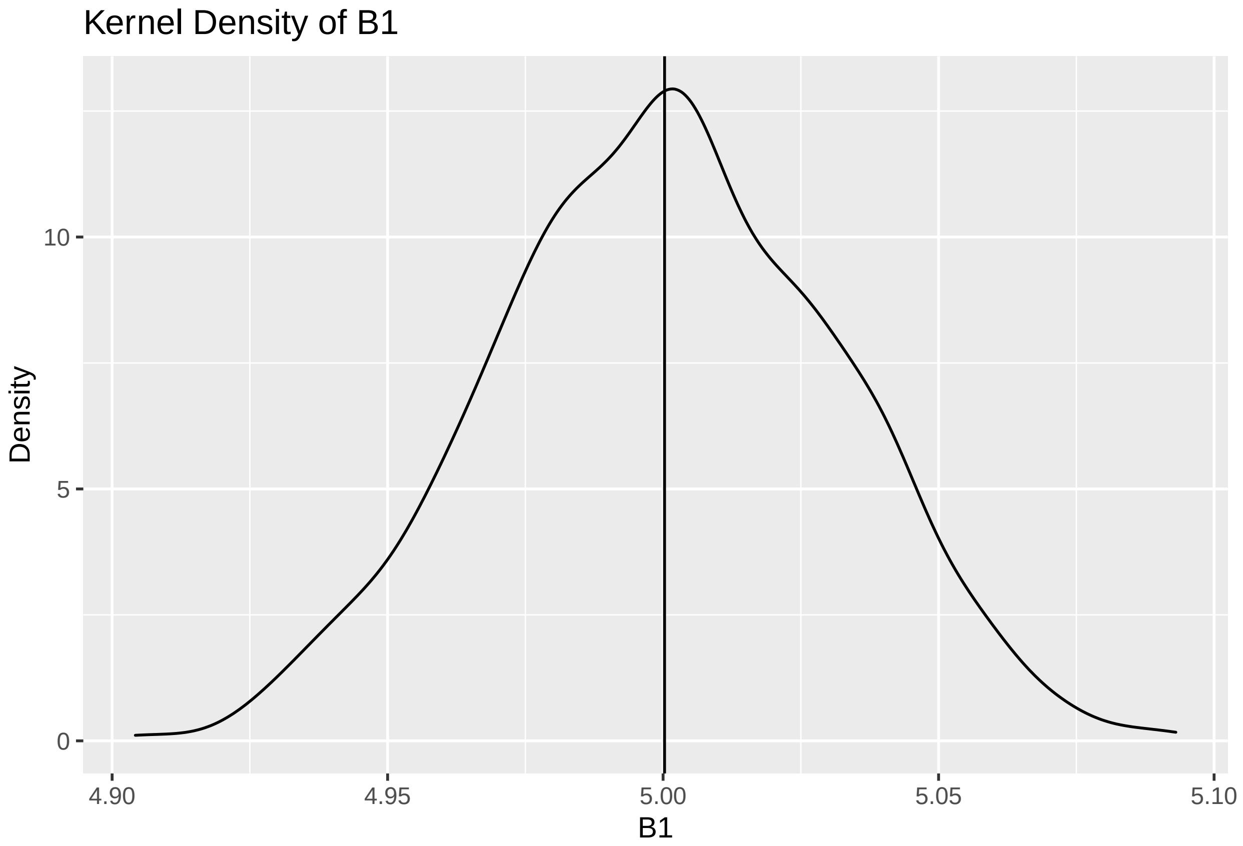
At the end of this process, you should have m draws of $\hat{\beta}_0$ and $\hat{\beta}_1$ which serves as draws from your sampling distributions. Generate a kernel density plot for your two sampling distributions. Superimpose a line on each for the mean of the distributions. From your simulations, does 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?

```
# Create empty dataframe

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`))

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

m <- summary(regression)

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

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

# I wanted to define a column as the standard error so I can better
# understand further questions that ask for it.

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

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



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

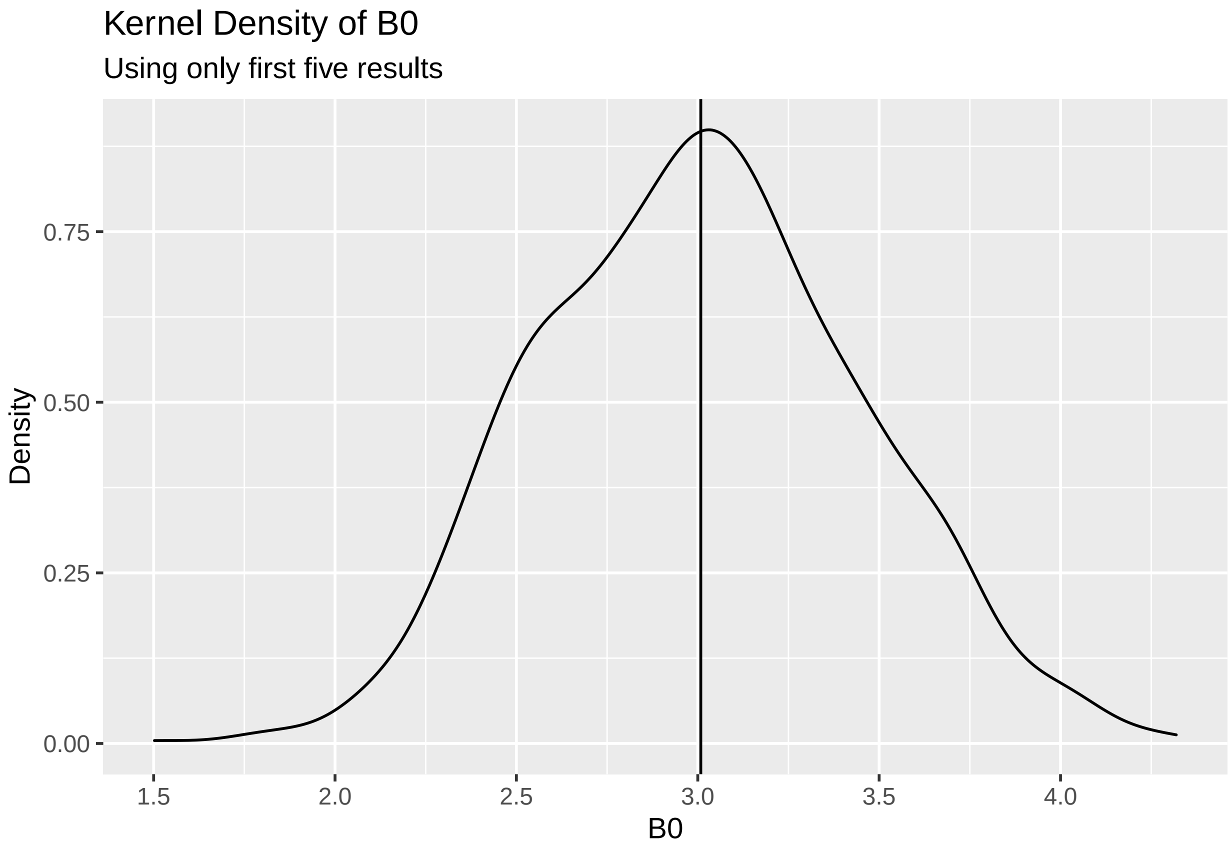


The density plots look fairly evenly spread/normally distributed around the sample means, which appear quite close to the original "3" and "5" coefficients (β_0 and β_1). So they do appear to be relatively unbiased. I also added a column "X3" to the dataframe (see my for loop) that pulls the standard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

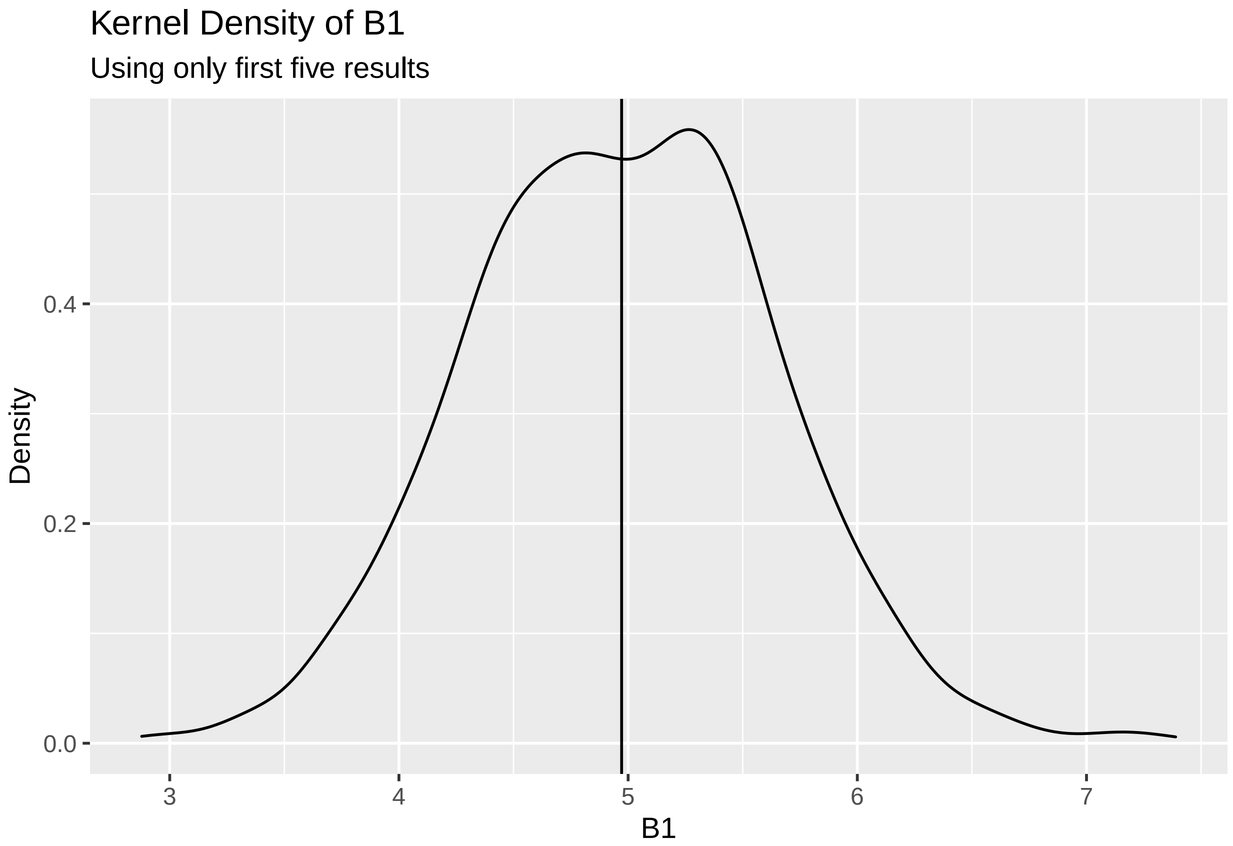
1b) Repeat (a), this time using just the first five observations of x ($n = 5$). How do your results compare? Why?

```
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]

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.

```
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?

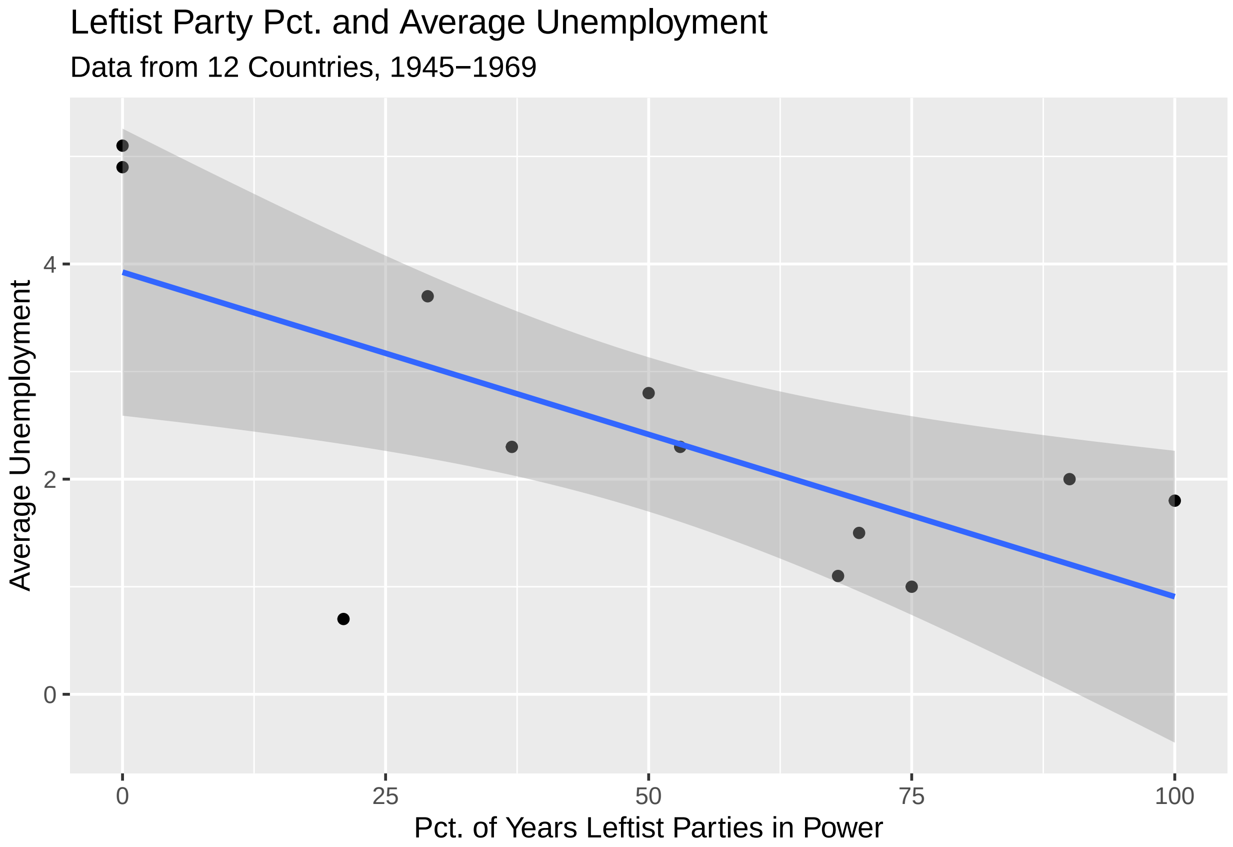
2a) Run a regression of unemployment on the independent variable government ideology.

```
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
```

```
ggplot(hibbs, aes(x = percleft, y = unemployment)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Pct. of Years Leftist Parties in Power",
       y = "Average Unemployment",
       title = "Leftist Party Pct. and Average Unemployment",
       subtitle = "Data from 12 Countries, 1945-1969")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Here, we see that there is a negative relationship between the percentage of years that leftist parties have been in power in a given country and the average unemployment rate in that country over that period of time. The results are significant at the 0.05 level. The R^2 value is 0.4657, implying that a bit under 50% of the variability of the dependent variable (unemployment) can be explained by the independent variable (percleft).

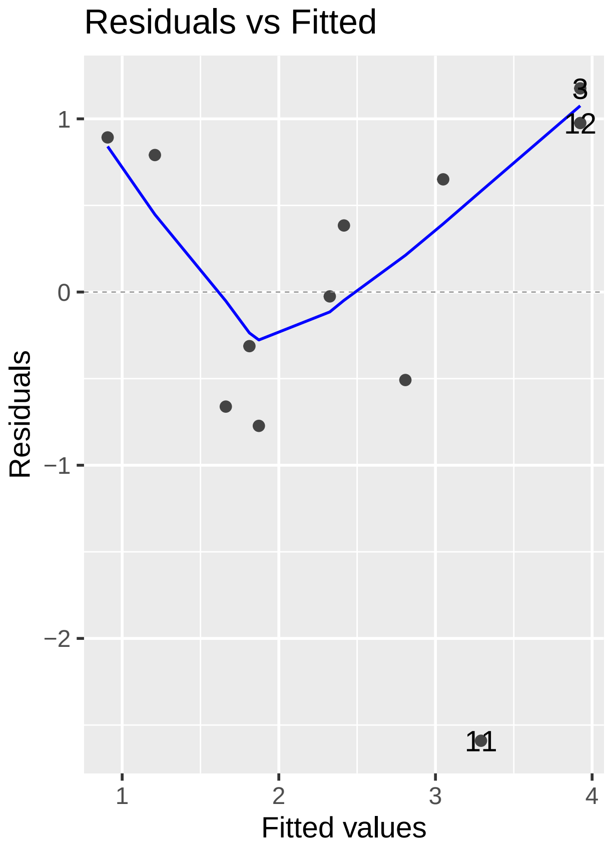
The R^2 value isn't bad, but I'm not entirely convinced that this model is practically useful. While the R^2 number implies that this variable certainly does have some explanatory power, I'm not convinced. The variable being explained (unemployment) is an incredibly complicated variable, and the amount of data is very small. I would be much more convinced by this high of an R^2 value if the model were bigger.

Let's look at some of the assumptions:

1) Linearity in Parameters:

A good test for linearity in parameters is the residuals vs. fitted values test. Below, we can run this:

```
autoplot(reg2a, 1)
```



It’s not great by any means, but at least the line is somewhat close to 0. I guess it could be worse, especially given the small sample size.

2) Random Sampling

This is definitely NOT a random sample. The author specifically chose 12 countries.

3) Variation in X

There is definitely variation; there are different values of x.

4) Zero conditional mean

In order to say that we have “zero conditional mean” it means that we have to have exogeneity. But here we definitely cannot discount the fact that the dependent variable could also depend on the independent variable. Maybe a certain rate of unemployment means that people are more likely to vote for more leftist governments. Also, my suspicion is that there are many omitted variables. We are only looking at a model with three variables here – percleft, inflation, and unemployment. These are all incredibly complex variables that presumably have many more explanations.

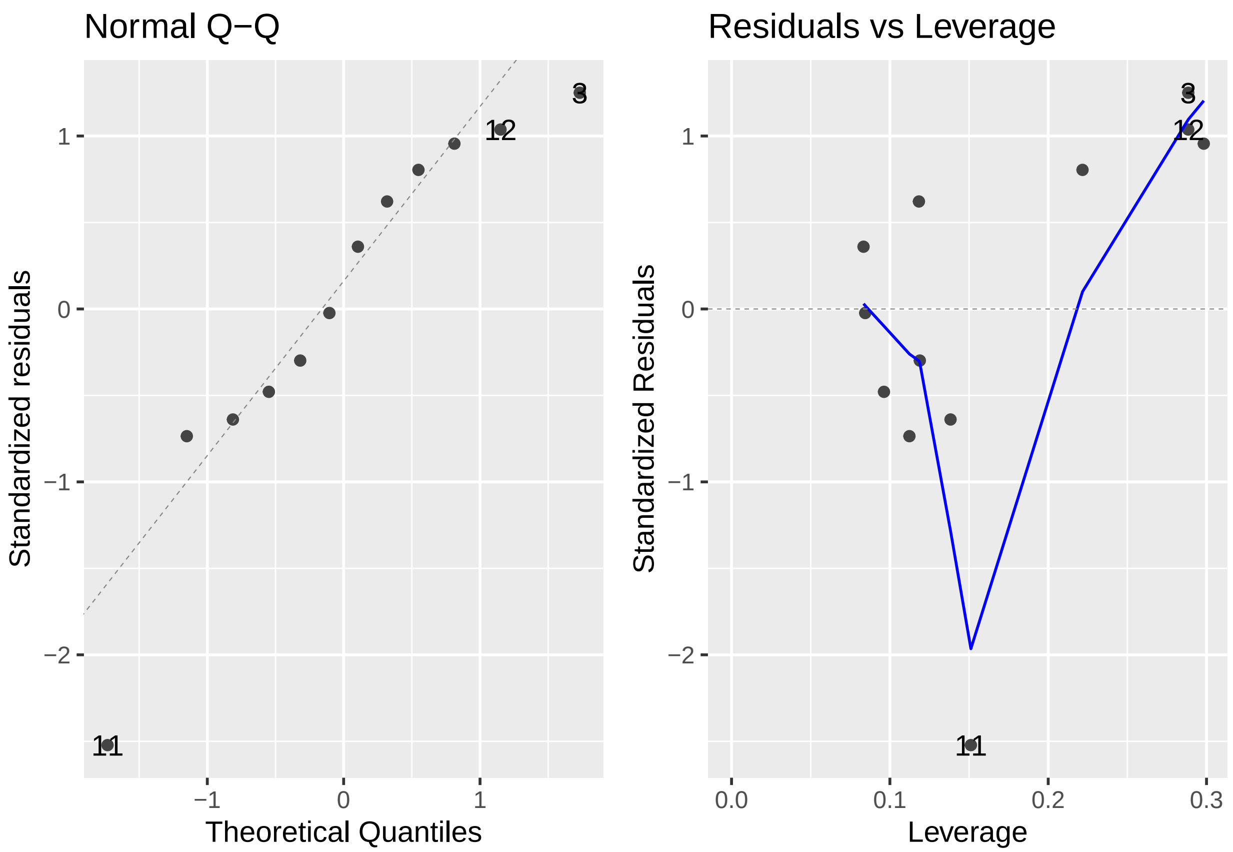
5) Homoskedasticity

Honestly, this looks fairly homoskedastic. I don’t see specific areas where there is more variation than others. There’s one outlier, but I don’t really feel like one outlier is enough to call a sample definitely non-homoskedastic.

6) Normality

Unfortunately, the errors are really not normally distributed. We can check by looking at these two plots. The outliers really make an impact on the regression.

```
autoplot(reg2a, c(2, 5))
```



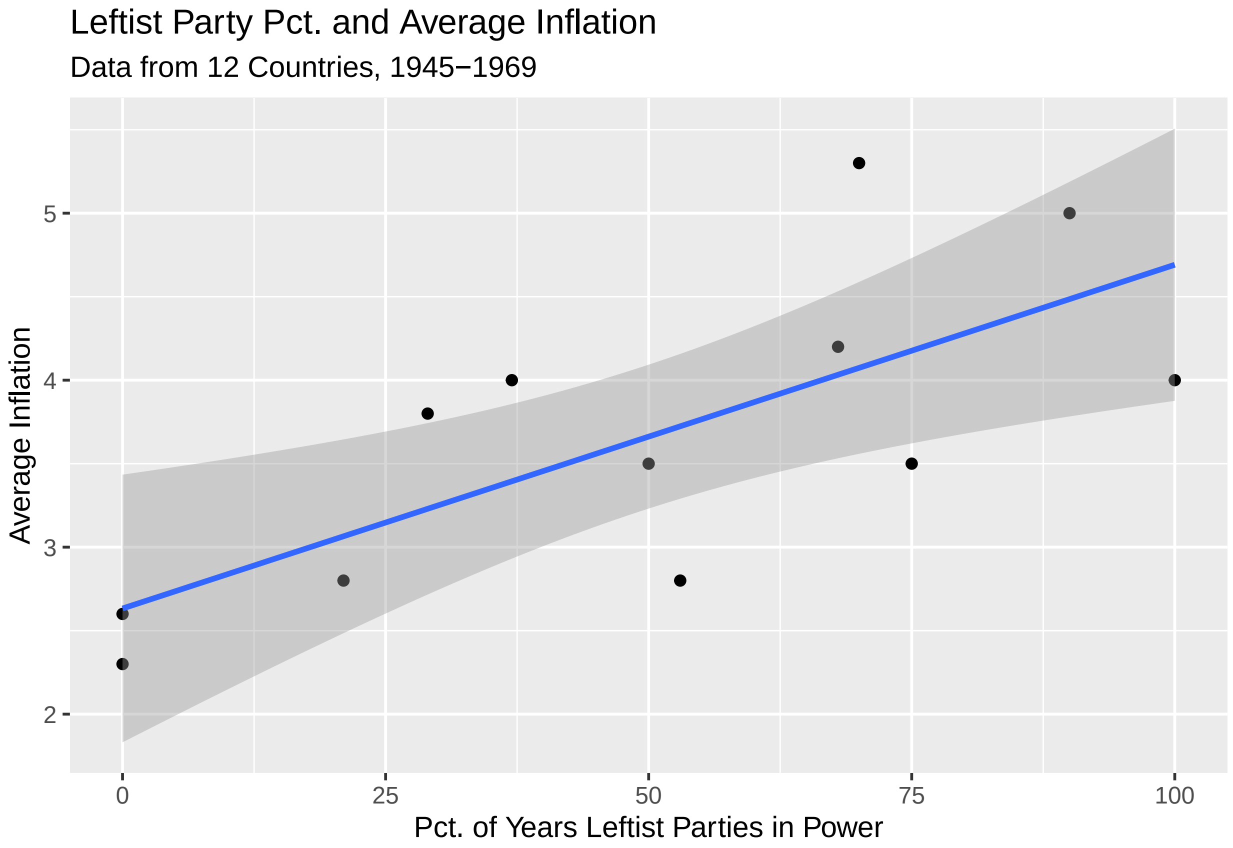
2b) Run a regression of inflation on the independent variable government ideology.

```
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
```

```
ggplot(hibbs, aes(x = percleft, y = inflation)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Pct. of Years Leftist Parties in Power",
       y = "Average Inflation",
       title = "Leftist Party Pct. and Average Inflation",
       subtitle = "Data from 12 Countries, 1945-1969")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Here, we see that there is a positive relationship between the percentage of years that leftist parties have been in power in a given country and the average inflation rate in that country over that period of time. The results are significant at the 0.01 level. The R^2 value is 0.5293, implying that a bit over 50% of the variability of the dependent variable (inflation) can be explained by the independent variable (percleft).

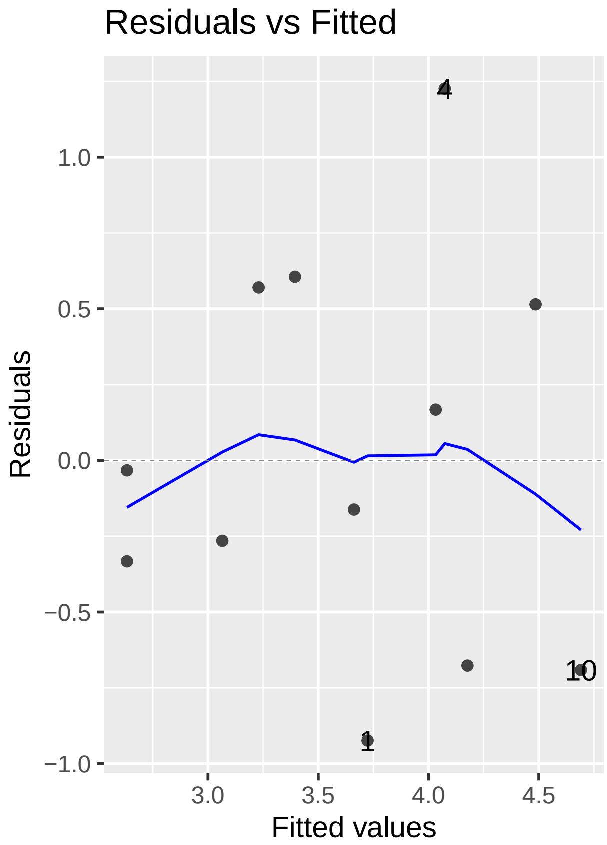
The R^2 value, again, isn’t bad, but the same concerns I had for the first regression apply here as well. Again, the variable being explained (inflation) is an incredibly complicated variable, and the amount of data is very small. I would be much more convinced by this high of an R^2 value if the model were bigger.

Let's look at some of the assumptions:

1) Linearity in Parameters:

A good test for linearity in parameters is the residuals vs. fitted values test. Below, we can run this:

```
autoplot(reg2b, 1)
```



This one is a lot better than the other one. The line is much straighter and is more centered around 0.

2) Random Sampling

This is definitely NOT a random sample. The author specifically chose 12 countries.

3) Variation in X

There is definitely variation; there are different values of x.

4) Zero conditional mean

In order to say that we have "zero conditional mean" it means that we have to have exogeneity. Again, however, we have the same issue as the previous regression. We cannot discount the fact that the dependent variable could also depend on the independent variable. Maybe a certain rate of inflation means that people are more likely to vote for more leftist governments. Also, my previously-expressed suspicion about likely omitted variable bias holds here as well.

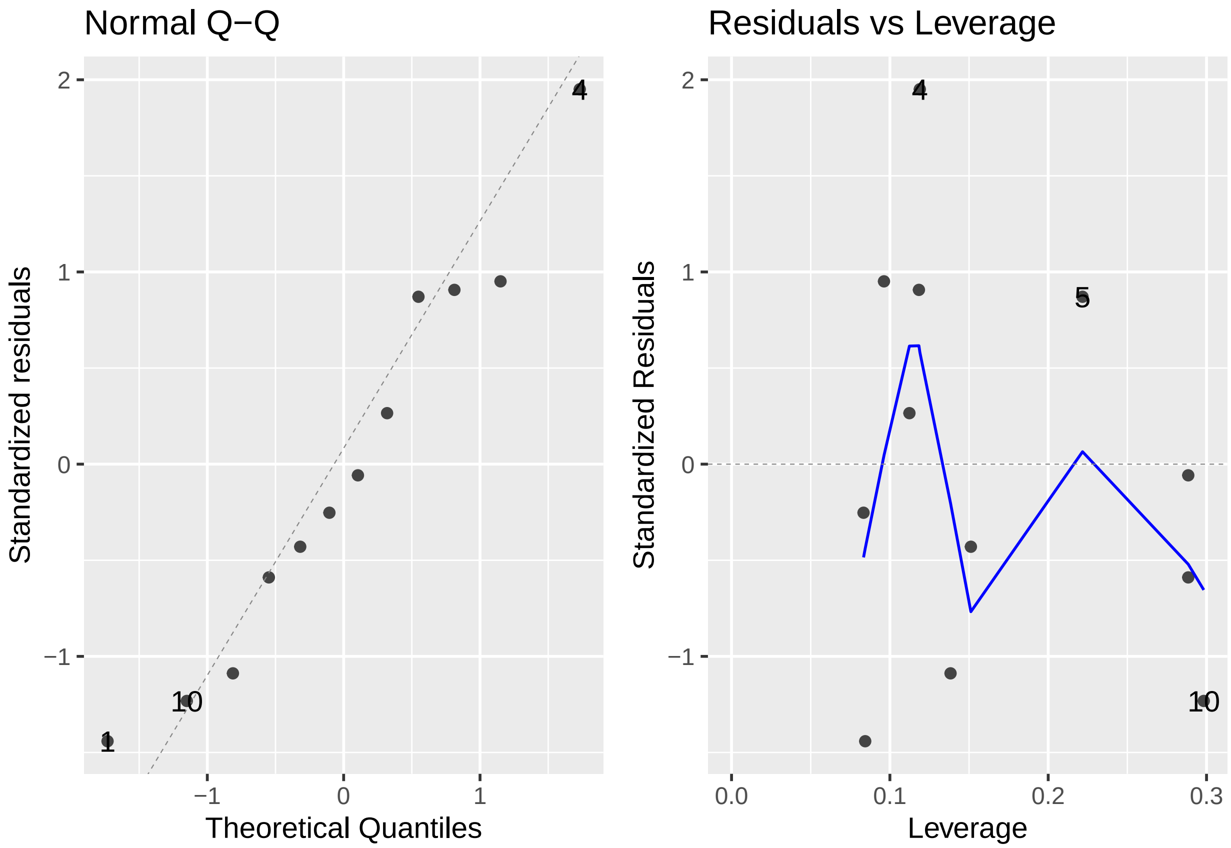
5) Homoskedasticity

This looks fairly homoskedastic as well. I don't see specific areas where there is more variation than others.

6) Normality

Let's do the same check as we did in 2a.

```
autoplot(reg2b, c(2, 5))
```



This relationship is not perfect, but is certainly much better than the other one. There is less jaggedness and there aren't quite as many outliers.

Question 3 Background:

Two weeks ago we looked at Olken's dataset, which showed that a random audit experiment did not reduce the amount of missing funds (a measure of corruption). A friend who hasn't taken PS 6800 comes to you with an observational dataset that they claim shows that audits do reduce corruption in a different case (let's say Argentina).

The dataset includes the variables `municipality`, `%missing`, and `audit`. `audit` is a binary variable taking 0 if a municipality did not conduct a corruption audit, and 1 otherwise and `%missing` is the proportion of funds that went missing (so it falls between 0 and 1). Municipal councils decide whether or not to conduct an audit. About 30% of observations for `%missing` are missing. Your friend used OLS to analyze `lm(%missing` $\sim$`audit)`. They find β_1 is equal to $-.15$, which they claim is the causal effect of audits on missing funds.

3a) Do you find the random sampling assumption for this dataset to be plausible? Why or why not?

I can't say for sure. I haven't seen the experiment structure yet – all we're looking at is a dataset. Plus, the 30% of missing values is a real red flag. I don't know the experimental design, I don't know how the sampling went, etc. I would say that it's theoretically possible, but I'd like to see more information.

3b) Do you find the zero conditional mean assumption for this model to be plausible? Why or why not?

I don't find the zero conditional mean assumption for this model to be plausible. There is a significant chance of endogeneity here; it could very well be that *because* a municipality is corrupt, they don't run audits. Thus, we can't fundamentally say that the zero conditional mean assumption is entirely plausible.

3c) Given your answers, do you believe that the municipal audits in Argentina lowered missing funds?

I cannot say that for sure. I think that the answers to the previous two questions throw a lot of doubt on this. For one, it is difficult to determine for sure which way the causal arrow runs, and also, I'm not sure how random the sample is.