Final Exam

Daniel Shapiro

12/19/2022

Question 1 Background:

The case Chavez v. Illinois State Police (1999) was a class-action lawsuit against the Illinois State Police for unconstitutional racial profiling. Among other evidence, the plaintiffs claimed that the percentage of drivers stopped who were Black was not representative of the proportion of the Illinois population that is Black.

The 1990 census found that the racial breakdown of Illinois is roughly 75% White and 15% Black (you can treat these statistics as fixed). The racial breakdown of stopped drivers was 68% White and 25% Black. The standard deviations for the estimated share of stopped drivers was 0.22 (White) and 0.17 (Black), in a sample of 50,000 stopped drivers.

1a) Imagine you are planning to statistically evaluate the plaintiffs' claims that Black drivers were much more likely to get pulled over relative to their population. Clearly specify a null and alternative hypothesis (5 points)

Null hypothesis (μ_0: Black drivers are not more likely to get pulled over relative to their population.

Alternative hypothesis (μ_a): Black drivers are more likely to get pulled over relative to their population.

1b) Calculate a test statistic that allows you to assess the amount of evidence against the claim that the stopped drivers represent a random sample from the Illinois population. Provide an approximate p-value. Explain, in words, what the test statistic and p-value mean in terms of substantive and statistical significance. (10 points)

The t statistic to test hypothesis μ_a against null hypothesis μ_0 can be calculated as: $\frac{\mu_a - \mu_0}{SE\widehat{(\mu_a)}}$

Here, we see that this is $\frac{.25-.15}{.17/\sqrt{(50000)}} = \frac{.1}{.0008} = 131.533$.

The t statistic, at first glance, looks abnormally large; after all, we are supposed to find out whether or not the t statistic fits between -1.96 and 1.96, and this number is nowhere close. But given the data, it makes sense: we have a massive sample size, so the level of precision of this sample will be quite high. The percentage of Black people stopped is **way** higher than would be expected if this sample were random, making it seem that there is almost no chance that this is a random sample. A high t statistic indicates large differences between the sample data and the null hypothesis, so this number is logical.

The p-value backs this assessment up. Using the pt() function, we see that the p-value is essentially 0. This number is most certainly below the significance level of $\alpha = 0.05$, so we can reject the null hypothesis at the 5% level of significance and say that the sample of stopped drivers does not represent a random sample from the Illinois population.

```
t <- -abs(131.533)
pt(q = t, df = 49999)
```

```
## [1] 0
```

```
# I only did one side (instead of multiplying it by 2) because we are only looking to
# see whether or not Black drivers are more likely to get pulled over.
```

1c) Calculate a confidence interval for your estimate of the difference. Interpret, in words, what this confidence interval means. Does this align with your findings in part b? (5 points)

The equation to find the confidence interval is: $\hat{\mu} \pm 1.96\hat{SE}$. Below, we can set this up.

```
muhat <- .25
se <- .17/sqrt(50000)

c((muhat - 1.96 * se), (muhat + 1.96 * se))
```

```
## [1] 0.2485099 0.2514901
```

This confidence interval shows us that based on the sample, we can expect that the real "percent Black" of drivers stopped in Illinois is between 24.85% and 25.15%, with 95% certainty. This does fit with our results for part b, as the two bounds are *nowhere close* to 15%, which would be the expected "percent Black" if Black people were indeed not more likely to get pulled over relative to their population, as stipulated by the null hypothesis.

Question 2 Background:

Imagine that the true population model between our variables of interest (Y and X_1) is causal only when controlling for X_2:

$$Y = -4 + 1.4X_1 - .7X_2 + \epsilon$$

Where:

$$\begin{aligned} X_1 &= N(5, 2) \\ X_2 &= .5X_1^2 - U(0, 4) \\ \epsilon &= N(0, 2) \end{aligned}$$

Unfortunately, a researcher has modeled the relationship between Y and X_1 using the following model:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

2) Draw a random sample from the population distribution of Y, X_1, and X_2 ($n = 1000$). How much does the researcher's model (once you estimate $\hat{\beta}_1$ and $\hat{\beta}_0$) depart from the true population model? Why? Can this be fixed?

First, I set up all of the parameters.

```r
dataframe <- data.frame(matrix(ncol = 2, nrow = 1000)) %>%
  mutate(X1 = rnorm(1000, mean = 5, sd = 2)) %>%
  mutate(X2 = runif(1000, min = 0, max = 4)) %>%
  mutate(epsilon = rnorm(1000, mean = 0, sd = 2))

for(i in 1:1000){
  dataframe$X2[i] <- .5 * (dataframe$X1[i]^2) - dataframe$X2[i]
  dataframe$Y[i] <- -4 + 1.4*dataframe$X1[i] - .7*dataframe$X2[i] + dataframe$epsilon[i]
}
```

Now, I run a regression using the model that the researcher used to see the discrepancy in coefficients.

```r
model2 <- lm(Y ~ X1, data = dataframe)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = dataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1301  -1.5844   0.3578   1.9495   9.6450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.97096    0.24338   16.32   <2e-16 ***
## X1          -1.97315    0.04545  -43.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.941 on 998 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6535
## F-statistic:  1885 on 1 and 998 DF,  p-value: < 2.2e-16
```

As we can see, the model that the researcher used is way wrong. The formula for the true model is $Y = -4 + 1.4X_1 - .7X_2 + \epsilon$, whereas this one ends up as about $Y = 4.85 - 2.13X_1 + \epsilon$.