Analyzing Data on Racial Prejudice in the 1980s

Eliza Durbin.

```r
#reading in data
listdata <- read.csv("~/Desktop/govt10-f21/Problem Sets/data/listexp.csv")
listdata <- na.omit(listdata)
```

```r
#average number of items selected between treatment and control group
t.test(listdata$y[listdata$treat==1], listdata$y[listdata$treat==0], #null hypothesis
       mu = 0, #90% confidence level
       conf.level=.90)
```

```
##
##  Welch Two Sample t-test
##
## data:  listdata$y[listdata$treat == 1] and listdata$y[listdata$treat == 0]
## t = 1.5139, df = 1315.9, p-value = 0.1303
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -0.006274468  0.150044313
## sample estimates:
## mean of x mean of y
##  2.190896  2.119011
```

The treatment group selected an average of 2.19 items, and the control group selected an average of 2.12 items. Therefore, we should fail to reject the null hypothesis. This makes sense because the t-value of 1.5139 is within the 90% confidence interval critical values of -1.64 and +1.64. Also, the confidence interval goes through zero, and the p-value is greater than .10. More proof that we cannot reject the null hypothesis at the 90% confidence interval.

```r
#separating southern and non-southern respondents
south<- subset(listdata, south==1)
non.south<- subset(listdata, south==0)

#repeating t-test
t.test(south$y[south$treat==1], south$y[south$treat==0], mu=0, conf.level=.90)
```

```
##
##  Welch Two Sample t-test
##
## data:  south$y[south$treat == 1] and south$y[south$treat == 0]
## t = 2.4471, df = 288.59, p-value = 0.015
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.08031924 0.41294135
## sample estimates:
```

```
## mean of x mean of y
##  2.203947  1.957317
```

```r
t.test(non.south$y[non.south$treat==1], non.south$y[non.south$treat==0], mu=0, conf.level=.90)
```

```
##
##  Welch Two Sample t-test
##
## data:  non.south$y[non.south$treat == 1] and non.south$y[non.south$treat == 0]
## t = 0.24602, df = 1009.1, p-value = 0.8057
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -0.07532091  0.10178594
## sample estimates:
## mean of x mean of y
##  2.187146  2.173913
```

For southern respondents, the average number of items selected in the treatment group is 2.20, and the average number of items selected in the control group is 1.96. Additionally, the confidence interval does not intercept zero, the p-value is less than .10, and the t-value is greater than the 90% interval critical values of 1.64. This means that we can reject the null hypothesis that there is no racial prejudice.

For non-southern respondents, the average number of items selected in the treatment group is 2.19, and the average number of items selected in the control group is 2.17. The confidence interval does go through zero, the t-value is much less than the critical values, and the p-value is greater than .10. Therefore, we fail to reject the null hypothesis that there is no racial discrimination.

```r
#changing confidence level for southern respondents
t.test(south$y[south$treat==1], south$y[south$treat==0], mu=0, conf.level=.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  south$y[south$treat == 1] and south$y[south$treat == 0]
## t = 2.4471, df = 288.59, p-value = 0.015
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -0.01470338  0.50796397
## sample estimates:
## mean of x mean of y
##  2.203947  1.957317
```

```r
1-.015
```

```
## [1] 0.985
```

The results do change. The 99% confidence interval goes through zero, so we fail to reject the null hypothesis at the 99% level. The p-value indicates that the maximum confidence in the estimate is 98.5% confidence. This is because we have confidence up until the point where the p-value is equal to alpha (which is one minus the p-value).

#Reading in new data set

```r
afb <- read.csv("~/Desktop/govt10-f21/Problem Sets/data/afb.csv")
```

```r
#regressing variables wifecoethnic with vote
result <- lm(data=afb,vote~wifecoethnic)

#identifying t and p-values
summary(result)
```

```
##
## Call:
## lm(formula = vote ~ wifecoethnic, data = afb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6077 -0.6077  0.3923  0.3923  0.6512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.607657   0.008239   73.75   <2e-16 ***
## wifecoethnic -0.258863   0.016931  -15.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4856 on 4550 degrees of freedom
## Multiple R-squared:  0.04886,    Adjusted R-squared:  0.04866
## F-statistic: 233.8 on 1 and 4550 DF,  p-value: < 2.2e-16
```

The t-value is 15.29 and very far outside of the critical values, and the p-value is very small. This means that we can reject the null hypothesis with 95% confidence. In other words, respondents who are coethnic with the president's wife are more likely to vote for the president.

```r
#creating a new binary variable in the dataset
afb$ethnicpercent2 <- ifelse(afb$ethnicpercent>median(afb$ethnicpercent), 1, 0)

#creating new subset of the data
afbsmall<- subset(afb, ethnicpercent2==0)
```

```r
#repeating regression with smaller dataset
result <- lm(data=afbsmall,vote~wifecoethnic)
summary(result)
```

```
##
## Call:
## lm(formula = vote ~ wifecoethnic, data = afbsmall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8881 -0.5990  0.4010  0.4010  0.4010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.59897    0.01041  57.539  < 2e-16 ***
```

```
## wifecoethnic  0.28914    0.04157   6.956 4.55e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4812 on 2278 degrees of freedom
## Multiple R-squared:  0.0208, Adjusted R-squared:  0.02037
## F-statistic: 48.39 on 1 and 2278 DF,  p-value: 4.553e-12
```

The conclusion does not change. We can still reject the null hypothesis. The t-value is still larger than the critical value and the p-value is very small.

```
#adding variable and regressing
result <- lm(data=afbsmall,vote~wifecoethnic + oppcoethnic)
summary(result)
```

```
##
## Call:
## lm(formula = vote ~ wifecoethnic + oppcoethnic, data = afbsmall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9086 -0.6000  0.4000  0.4000  0.4318
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.60003    0.01053  56.975  < 2e-16 ***
## wifecoethnic  0.30859    0.05078   6.077 1.43e-09 ***
## oppcoethnic  -0.03187    0.04778  -0.667    0.505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4813 on 2277 degrees of freedom
## Multiple R-squared:  0.02099,    Adjusted R-squared:  0.02013
## F-statistic: 24.41 on 2 and 2277 DF,  p-value: 3.242e-11
```

The ethnicity of the opponent is not a major confounder. Even with this variable, the wifecoethnic variable is still statistically significant at the 99.9% level– with a small p-value and large t-value. Also, the coefficient for wifecoethnic –which shows that, holding all other variables constant, the respondent being the same ethnicity as the president's wife increases vote probability by around 30.9%– does not change very much when adding the oppcoethnic variable. This means that wifecoethnic explains most of the vote probability. Also, holding all other variables constant, the respondent being the same race as the presidential opponent decreases vote probability by 3.19%. In other words, this variable is not statistically significant (further seen through its large p-value).