HW7_Markdown

2024-03-20

Manay Divatia

md46245

1

a)

```
survey_data = read.csv("~/Downloads/finlit15and18.csv")
library(boot)

num_observations <- nrow(survey_data)

gender_breakdown <- table(survey_data$A3)
age_breakdown <- summary(survey_data$A3A)

income_breakdown <- table(survey_data$A8)
```

There are 4694 observations in this dataset. The gender breakdown is 2539 males and 2155 females which is a difference of 384. It does seem roughly even which means the data for both will be a good estimate of the population. The age breakdown is between 18 and 86 years old. The household income ranged from 1-8 with 6 and 7 being the most common.

b)

```
gender_grouped <- split(survey_data$literacy, survey_data$A3)
mean_literacy_female <- mean(gender_grouped$"1")
mean_literacy_male <- mean(gender_grouped$"2")
observed_difference <- mean_literacy_female - mean_literacy_male

diff_mean <- function(data, indices) {
  female_mean <- mean(data[indices[1]][data$A3 == "Female"])
  male_mean <- mean(data[indices[2]][data$A3 == "Male"])
  return(female_mean - male_mean)
}
```

The average literacy difference between females and males is 0.5993832. It seems that there is not a significant different from zero so we can't draw any conclusions.

c)

```
fit <- lm(literacy ~ factor(A3), data = survey_data)
summary(fit)
```

```
##
## Call:
## lm(formula = literacy ~ factor(A3), data = survey_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1363 -1.1363 -0.1363  0.8637  2.4631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.13627    0.02841   145.6   <2e-16 ***
## factor(A3)2 -0.59938    0.04192   -14.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.431 on 4692 degrees of freedom
## Multiple R-squared:  0.04175,    Adjusted R-squared:  0.04154
## F-statistic: 204.4 on 1 and 4692 DF,  p-value: < 2.2e-16
```

```
coef_boot <- function(data, indices) {
  fit <- lm(literacy ~ factor(A3), data = data[indices, ])
  return(coef(fit)[2])
}

boot_result_coef <- boot(survey_data, coef_boot, R = 1000)
sampling_std <- sd(boot_result_coef$t)
regression_se <- summary(fit)$coefficients[2, "Std. Error"]
```

The standard deviation of the sampling distribution is 0.043 while the standard error for the regression output is 0.042 which is very similar in this case.

d)

```
small_model <- lm(Y ~ literacy + A8, data = survey_data)
summary(small_model)
```

```
##
## Call:
## lm(formula = Y ~ literacy + A8, data = survey_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04275 -0.71992  0.06955  0.78400  2.51859
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.418176   0.072641 -19.523  < 2e-16 ***
## literacy     0.072212   0.009965   7.247 4.97e-13 ***
## A8           0.181250   0.011361  15.953  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9601 on 4691 degrees of freedom
## Multiple R-squared:  0.07857,    Adjusted R-squared:  0.07817
## F-statistic:   200 on 2 and 4691 DF,  p-value: < 2.2e-16
```

```r
large_model <- lm(Y ~ literacy + A5_2015 + A3A + J2 + A3 + A8 + E20 + F2_2 + F2_3 + F2_4 + F2_5 + F2_6,
summary(large_model)
```

```
##
## Call:
## lm(formula = Y ~ literacy + A5_2015 + A3A + J2 + A3 + A8 + E20 +
##     F2_2 + F2_3 + F2_4 + F2_5 + F2_6, data = survey_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89583 -0.53471  0.05947  0.57845  2.84092
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.763e+00  1.238e-01 -30.405  < 2e-16 ***
## literacy    -3.651e-02  9.290e-03  -3.930 8.61e-05 ***
## A5_2015     -1.757e-02  8.406e-03  -2.090   0.0367 *
## A3A          8.828e-05  1.249e-03   0.071   0.9436
## J2           5.788e-02  5.608e-03  10.321  < 2e-16 ***
## A3          -3.874e-02  2.504e-02  -1.547   0.1219
## A8           1.020e-01  1.010e-02  10.102  < 2e-16 ***
## E20          3.947e-01  4.034e-02   9.783  < 2e-16 ***
## F2_2         5.082e-01  2.689e-02  18.899  < 2e-16 ***
## F2_3         5.535e-01  3.017e-02  18.349  < 2e-16 ***
## F2_4         3.388e-01  3.742e-02   9.054  < 2e-16 ***
## F2_5         1.238e-01  5.129e-02   2.414   0.0158 *
## F2_6        -9.593e-02  4.469e-02  -2.147   0.0319 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7952 on 4681 degrees of freedom
## Multiple R-squared:  0.3693, Adjusted R-squared:  0.3677
## F-statistic: 228.4 on 12 and 4681 DF,  p-value: < 2.2e-16
```

```r
literacy_effect_small <- coef(small_model)["literacy"]
literacy_effect_large <- coef(large_model)["literacy"]

rsquared_small <- summary(small_model)$r.squared
rsquared_large <- summary(large_model)$r.squared
```

I think the small model is good at just focusing on the variables that are important. What we saw with the small model is that the r squared value was .0785 while for the large model is was much higher at .37. The large model was able to utilize all the data to lower variation in predict and expected value. However, there were variables that did not meet the 0.05 requirement so we couldn't say that they were significant and removing those would help us lower the difference between predicted and expected value even more.

2

a)

```r
transfer_data = read.csv("~/Downloads/transfer.csv")
cutoffs <- c(10188, 13584, 16980)

transfer_data$closest_cutoff <- sapply(transfer_data$pop82, function(x) min(abs(x - cutoffs)))
```

```r
transfer_data$normalized_percent_score <- (transfer_data$closest_cutoff / cutoffs) * 100
```

```
## Warning in transfer_data$closest_cutoff/cutoffs: longer object length is not a
## multiple of shorter object length
```

b)

```r
subset_data <- transfer_data[transfer_data$normalized_percent_score <= 3, ]

model_poverty <- lm(poverty91 - poverty80 ~ 1, data = subset_data)
model_educ <- lm(educ91 - educ80 ~ 1, data = subset_data)

summary(model_poverty)
```

```
##
## Call:
## lm(formula = poverty91 - poverty80 ~ 1, data = subset_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26240 -0.05473 -0.00798  0.05245  0.31987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.005814   0.004877   1.192    0.234
##
## Residual standard error: 0.08587 on 309 degrees of freedom
```

```r
summary(model_educ)
```

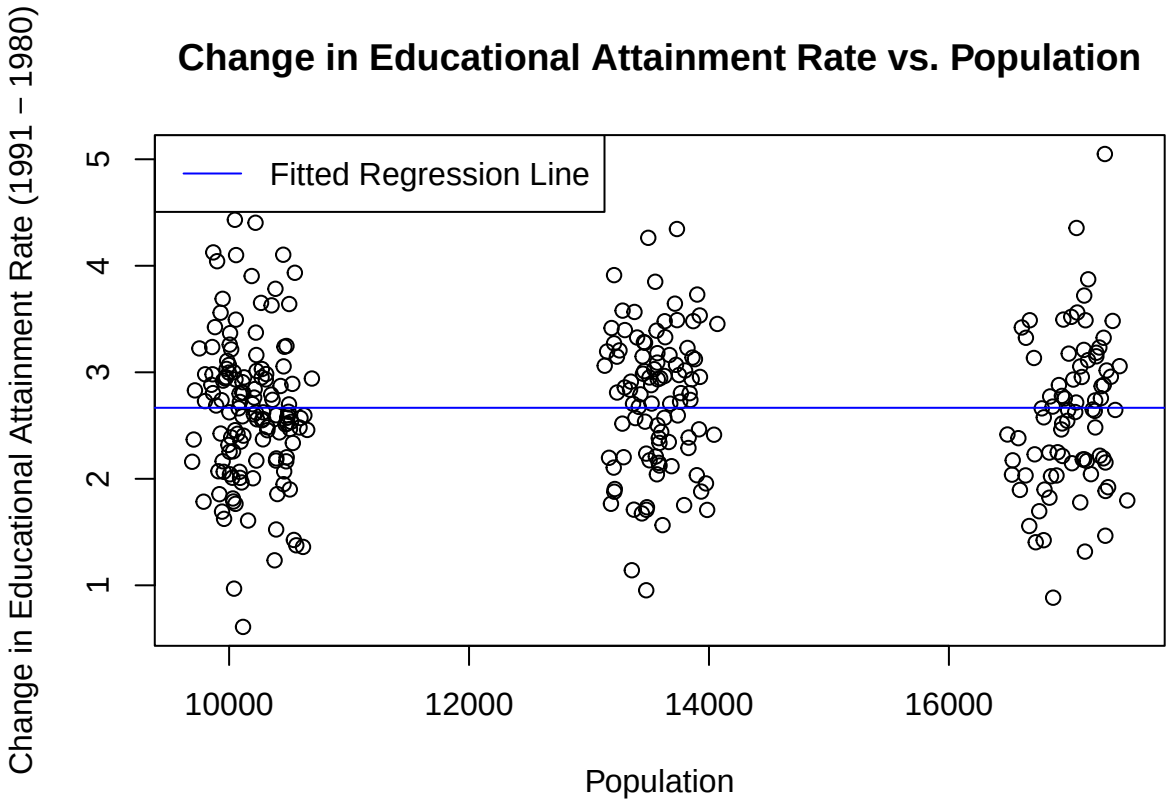
```
##
## Call:
## lm(formula = educ91 - educ80 ~ 1, data = subset_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05726 -0.48173  0.01617  0.40293  2.38267
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.66668    0.03874   68.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6821 on 309 degrees of freedom
```

For the average causal effect of government transfer on poverty, I used a linear model to do this. I found that the median was -0.00798 and the residual standard error was 0.085. Because the median in difference in poverty was so close to zero and the first and third quartile contained zero, it seems like we can't make any claims based on this data. I did the same thing for the literacy and educational attainment variables. For both, I found there to be a similar case with the median in difference in educational attainment to be 0.016 with an estimated standard eror of 2.66.

c)

```
plot(subset_data$pop82, subset_data$educ91 - subset_data$educ80, xlab = "Population", ylab = "Change in
abline(model_poverty, col = "blue")
legend("topleft", legend = c("Fitted Regression Line", "Population Threshold"), col = c("blue", "red"),
```


```
plot(subset_data$pop82, subset_data$educ91 - subset_data$educ80, xlab = "Population", ylab = "Change in
abline(model_educ, col = "blue")
legend("topleft", legend = c("Fitted Regression Line"), col = c("blue"), lty = 1:2)
```



The plot backs up the data because we can see about half the points above the blue line and half below for each cutoff which suggests that there wasn't a significant difference for both the poverty model and the educational attainment model.

d)

```
subset_data$above_threshold <- ifelse(subset_data$pop82 > subset_data$closest_cutoff, 1, 0)

mean_diff_educ <- tapply(subset_data$educ91, subset_data$above_threshold, mean)
mean_diff_literate <- tapply(subset_data$literate91, subset_data$above_threshold, mean)
mean_diff_poverty <- tapply(subset_data$poverty91, subset_data$above_threshold, mean)
```

The mean difference in educational attainment rate was the highest at 4.54. The mean difference for literacy and poverty were similar where the mean difference in literacy was .79 and the mean difference in povery was .61. I feel that the estimates are more appropriate in part d because ultimately we are finding the difference in means and that seems more valuable than using a linear regression model.