Hypothesis Testing of Adult Demographics Dataset

2024-10-12

Import libraries first

```r
library(ggplot2)
library(dplyr)
library(rcompanion)
library(tidyr)
library(patchwork)
```

Prepare dataset

```r
df <- read.csv("adult_data.csv", header = TRUE, sep = ",")
summary(df)
```

```
##       age         workclass             fnlwgt         education        
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561      
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character  
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character  
##  Mean   :38.58                      Mean   : 189778                     
##  3rd Qu.:48.00                      3rd Qu.: 237051                     
##  Max.   :90.00                      Max.   :1484705                     
##  education.num   marital.status      occupation        relationship      
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561      
##  1st Qu.: 9.00   Class :character   Class :character   Class :character  
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :10.08                                                            
##  3rd Qu.:12.00                                                            
##  Max.   :16.00                                                            
##      race               sex             capital.gain    capital.loss   
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0  
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0  
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0  
##                                        Mean   : 1078   Mean   :  87.3  
##                                        3rd Qu.:    0   3rd Qu.:   0.0  
##                                        Max.   :99999   Max.   :4356.0  
##  hours.per.week  native.country        salary         
##  Min.   : 1.00   Length:32561       Length:32561      
##  1st Qu.:40.00   Class :character   Class :character  
##  Median :40.00   Mode  :character   Mode  :character  
##  Mean   :40.44                                        
##  3rd Qu.:45.00                                        
##  Max.   :99.00                                        
```

Questions (Re-hashed from SQL analysis)

The questions posited earlier for the SQL analysis were

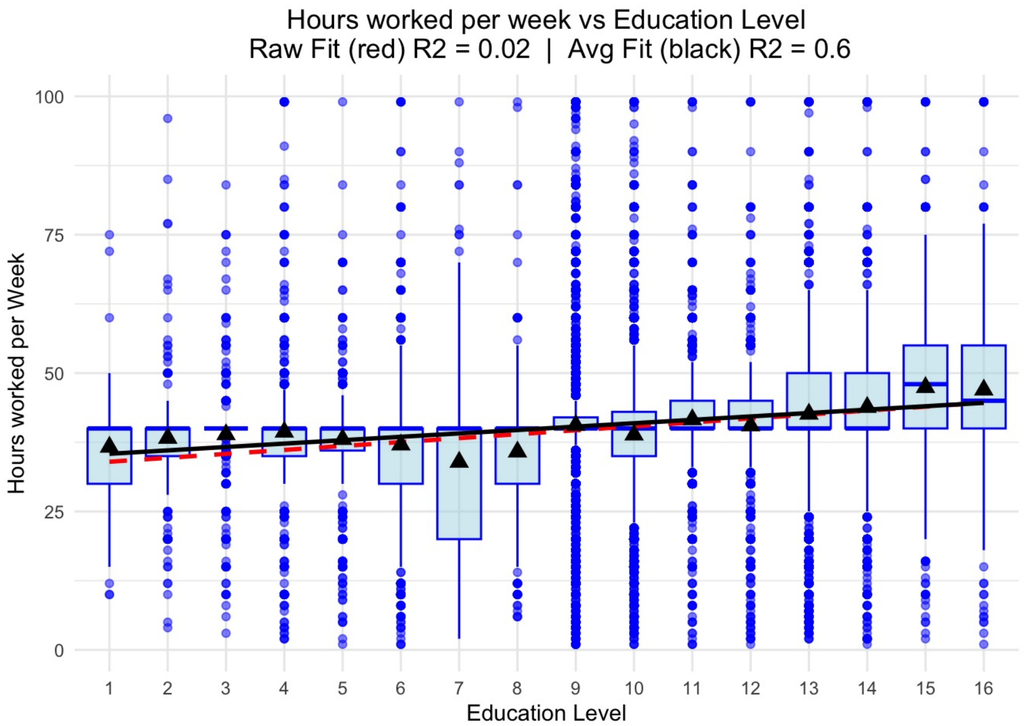
1. What’s the relationship between hours-per-week worked and education level?
2. Is there a difference in education level between U.S. born and non-U.S. born persons?
3. Does sex affect the likelihood of earning greater than $50k in salary?
4. For those earning a salary >$50k, is there a change in hours-per-week worked between different racial categories?

We’ll now try to get statistical answers to these questions, with quantification when possible.

Q1: What’s the relationship between hours-per-week worked and education level?

Let’s first visualize the hours worked per week by education level. We’ll show the data for each education level as a boxplot and a linear fit for both the hours worked per week as well as the *average* hours worked per week by education level. The average number of hours worked per week is shown with a black triangle for each level.

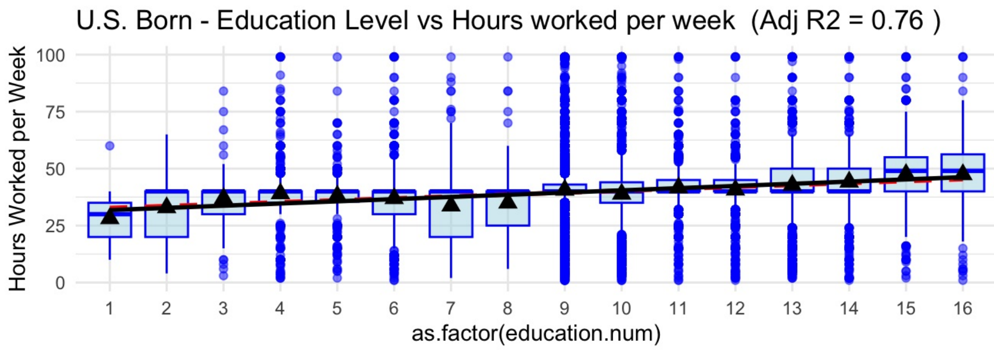
```
# get avgs hours worked per week by education level
df_avg <- df %>%
  group_by(education.num) %>%
  summarise(avg_hours_per_week = mean(hours.per.week))

# make fits for hours worked per week and avg hours worked
model_total <- lm(hours.per.week ~ education.num, data = df)
model_avg <- lm(avg_hours_per_week ~ education.num, data = df_avg)

# Create the plot
ggplot(df, aes(x = as.factor(education.num), y = hours.per.week)) +
    geom_boxplot(alpha = 0.5, color = "blue", fill = "lightblue") +
    geom_smooth(aes(group = 1), method = "lm", se = FALSE, color = "red", linetype = "dashed") +
    geom_smooth(data = df_avg,
                aes(x = as.factor(education.num),
                y = avg_hours_per_week, group = 1),
                method = "lm", se = FALSE, color = "black") +
    geom_point(data = df_avg, aes(x = as.factor(education.num), y = avg_hours_per_week),
               color = "black", size = 3, shape = 17) +
    labs(title = paste("Hours worked per week vs Education Level\n",
                       "Raw Fit (red) R2 =", round( summary(model_total)$r.squared, 2 )," | ",
                       "Avg Fit (black) R2 =", round( summary(model_avg)$r.squared, 2 )),
         x = "Education Level",
         y = "Hours worked per Week") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))
```

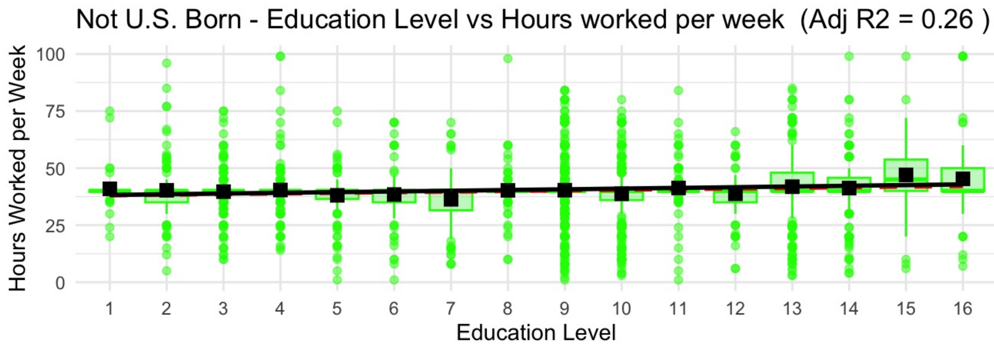


There's clearly far, far too much variance in each education level to be able to estimate the hours worked per week (the R squared for that fit, the red line above, is about 0.02!) - using ANOVA with a post-hoc Tukey test is not in the cards here because there's no way the residuals would by normally distributed or homoskedastic. Instead, we can model the average amount of hours worked per week by education level instead, which is a lot more reasonable with an R squared of about 0.6). There is some non-linearity present at education levels 8 and below though. If we recall from the SQL analysis, there's was quite a disparity between persons born in the U.S. and those not by education level. It's also well known that the United States relies heavily on non-citizens for "off the books" labor which employers exploit or order to reduce their labor costs. This means that U.S. born status affects both the independent and dependent variables here and acts as a confounder. Let's look at these distributions separately and see if the fit improves.

```r
# create boolean variable with True if U.S. born and False otherwise
df <- df %>%
  mutate(us_born = native.country == 'United-States')

# make separates dfs for U.S. born status
df_us_born <- df %>%
  filter(us_born == TRUE)
df_non_us_born <- df %>%
  filter(us_born == FALSE)

# Compute avg hours worked per week per educaiton level for both groups
df_avg_us <- df_us_born %>%
  group_by(education.num) %>%
  summarise(avg_hours_per_week = mean(hours.per.week))
df_avg_non_us <- df_non_us_born %>%
  group_by(education.num) %>%
  summarise(avg_hours_per_week = mean(hours.per.week))

# make linear fits over averages
model_us <- lm(avg_hours_per_week ~ education.num, data = df_avg_us)
model_non_us <- lm(avg_hours_per_week ~ education.num, data = df_avg_non_us)

# U.S. born plot
plot_us_born <- ggplot(df_us_born, aes(x = as.factor(education.num), y = hours.per.week)) +
    geom_boxplot(alpha = 0.5, color = "blue", fill = "lightblue") +
    geom_smooth(aes(group = 1), method = "lm", se = FALSE, color = "red", linetype = "dashed") +
    geom_smooth(data = df_avg_us,
                aes(x = as.factor(education.num),
                    y = avg_hours_per_week,
                    group = 1),
                method = "lm", se = FALSE, color = "black") +
    geom_point(data = df_avg_us, aes(x = as.factor(education.num), y = avg_hours_per_week),
               color = "black", size = 3, shape = 17) +
    labs(title = paste("U.S. Born - Education Level vs Hours worked per week ",
                       "(Adj R2 =", round(summary(model_us)$adj.r.squared, 2), ")"),
         y = "Hours Worked per Week") +
    theme_minimal()

# Non-U.S. born plot
plot_non_us_born <- ggplot(df_non_us_born, aes(x = as.factor(education.num), y = hours.per.week)) +
    geom_boxplot(alpha = 0.5, color = "green", fill = "lightgreen") +
    # Linear fit for raw data
    geom_smooth(aes(group = 1), method = "lm", se = FALSE, color = "red", linetype = "dashed") +
    # Linear fit for averages
    geom_smooth(data = df_avg_non_us, aes(x = as.factor(education.num), y = avg_hours_per_week, group = 1),
                method = "lm", se = FALSE, color = "black") +
    geom_point(data = df_avg_non_us, aes(x = as.factor(education.num), y = avg_hours_per_week),
               color = "black", size = 3, shape = 15) +
    labs(title = paste("Not U.S. Born - Education Level vs Hours worked per week ",
                       "(Adj R2 =", round(summary(model_non_us)$adj.r.squared, 2), ")"),
         x = "Education Level",
         y = "Hours Worked per Week") +
    theme_minimal()

# Combine the two plots into one figure with two rows
plot_us_born / plot_non_us_born
```

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






```
# show summary for linear fits
print(summary(model_us))
```

```
##
## Call:
## lm(formula = avg_hours_per_week ~ education.num, data = df_avg_us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8862 -1.6232  0.1853  1.4456  4.3762
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.8378     1.3339   23.12 1.50e-12 ***
## education.num   0.9615     0.1379    6.97 6.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.544 on 14 degrees of freedom
## Multiple R-squared:  0.7763, Adjusted R-squared:  0.7603
## F-statistic: 48.59 on 1 and 14 DF,  p-value: 6.548e-06
```

```
print(summary(model_non_us))
```

```
##
## Call:
## lm(formula = avg_hours_per_week ~ education.num, data = df_avg_non_us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7570 -1.3678 -0.1315  1.3642  4.4664
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    37.9288     1.1799  32.145 1.61e-14 ***
## education.num   0.3088     0.1220   2.531    0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.25 on 14 degrees of freedom
## Multiple R-squared:  0.3138, Adjusted R-squared:  0.2648
## F-statistic: 6.404 on 1 and 14 DF,  p-value: 0.02401
```

It's pretty clear that the linear fit improves for those born in the U.S. Given an adjusted R squared of 0.76 and a small p-value on the 'education.num' coefficient, we can reasonably interpret the model to state that an increase in education level for U.S. born persons results in about 1 additional hour worked per week *on average*.

The fit for those not born in the U.S. is quite poor with an adjusted R squared of 0.31. Again, the p-value associated with the coefficient for education level is significant, but only somewhat. This suggests that there is a relationship between the average number of hours worked per week and education level for those not born in the U.S., but it is not sufficiently explained in this model alone. One possible explanation is that this data is comprised of two subgroups: those not born in the U.S. who come to the United States with little education for work as well as those who come to the United States for work after higher education. Separating these groups might actually give a more insightful analysis, but that's beyond the scope of what's being done here.

Q1 Answer: The hours worked per week possesses too much variance to be able to explain via education level, however the average number of hours is more reliable. For those born in the United States, there is 1 additional hour of work per week on average associated with increasing education level. For non U.S. born individuals, while there is a statistically significant relationship between education level and hours worked, the linear model explains much less of the variation, indicating that other factors likely contribute to the complexity of this relationship.

Q2: Is there a difference in education level between U.S. born and non-U.S. born persons?

This question was partially addressed previously, but we can perform a separate hypothesis test to determine the significance between education level and U.S. born status directly. We'll perform a simple Chi-square independence test.

```
# chi square test
contingency_table <- table(df$education.num, df$us_born)
chisq_test <- chisq.test(contingency_table)
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 2618.1, df = 15, p-value < 2.2e-16
```

As expected from the SQL analysis, there's a statistically significant difference in education for U.S. born persons vs non-U.S. born from the p-value. This only tells us that a difference exists, but doesn't highlight which education levels these correspond to. We can investigate the standardized residuals from the chi-square test and see which levels have magnitude greater than 2 (meaning, beyond two standard deviations).

```
print(chisq_test$stdres)
```

```
##
##             FALSE       TRUE
##   1    13.162303 -13.162303
##   2    26.464707 -26.464707
##   3    36.304750 -36.304750
##   4    10.372605 -10.372605
##   5     9.529825  -9.529825
##   6    -1.323015   1.323015
##   7    -1.397734   1.397734
##   8     3.628102  -3.628102
##   9   -11.435031  11.435031
##   10   -9.066116   9.066116
##   11   -4.583167   4.583167
##   12   -2.661957   2.661957
##   13    1.532656  -1.532656
##   14    1.342245  -1.342245
##   15    1.928772  -1.928772
##   16    6.807653  -6.807653
```

The levels with greater than 2 are 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, and 16, which roughly corrobarate the earlier observations, repeated here below:

1. "The Doctorate (education-num = 16) education level has twice the relative representation among non-U.S. born persons than those born in the U.S" in reference to level 16.
2. "The Prof-school, Masters, and Bachelors (15 - 13) levels show more even representation." Levels 13 - 15 were all within 2 sigma.
3. "Those U.S. born with Associates, some college, or complete high school (12 - 9) have greater representation." Levels 9 - 12 all had greater than 2 sigma.
4. "10th through 12th grade high school education (6 - 8) has somewhat even representation." This is somewhat correct as levels 6 and 7 were not statistically significant, but 8 was.
5. "Non-U.S. born persons with an education level below 10th grade high school (<= 5) are much more populous than U.S. born." Levels 1 - 5 possessed some of the largest standardized residuals, indicating the largest deviations.

Q2 Answer: There is a statistically significent difference between U.S. born and non-U.S. born persons at most education levels. Particularly, it's apparent that:

1. Low education levels (1-5) have a much higher representation among non-U.S. born persons.

2. Higher education levels (e.g., Doctorate) have a greater relative representation among non-U.S. born persons.

3. Intermediate levels (e.g., Associate's, high school diploma) show greater representation among U.S. born persons.

Q3: Does sex affect the likelihood of earning greater than $50k in salary?

The data for this question is just a simple 2 x 2 table problem, which can be answered with either a propotion test or a chi-square indpendence test. A chi-square test was used to answer Q2, so we'll opt for the proportion test here instead (and also since we can get a difference in proportions from it). Calling the test is pretty simple and requires little wrangling:

```
sex_salary_table <- table(df$sex, df$salary)
print(sex_salary_table)
```

```
##
##          <=50K  >50K
##   Female  9592  1179
##   Male   15128  6662
```

```
prop_result <- prop.test( sex_salary_table )
print(prop_result)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  sex_salary_table
## X-squared = 1517.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1877104 0.2048416
## sample estimates:
##    prop 1    prop 2
## 0.8905394 0.6942634
```

These results are statistically significant given the p-value and mean that 89% (= 9592 / (9592 + 1179)) of females earn <=$50k in salary [prop 1] while 69% (= 15128 / (15128 + 6662)) of males earn <=$50k in salary [prop 2]. The 95% confidence interval indicates that males are 18.8% to 20.4% to earn >$50k in salary than females.

Q3 Answer: Males are statistically significantly more likely than females to earn more than $50k. Specifically, males are 18.8% to 20.4% more likely than females to earn >$50k in salary.

Q4: For those earning a salary >$50k, is there a change in hours-per-week worked between different racial categories?

We'll run an ANOVA test to compare the hours-per-week distributions between different racial categories.

```
df_over_50k <- df[df$salary == ">50K", ]
anova_result <- aov(hours.per.week ~ race, data = df_over_50k)
summary(anova_result)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race           4    650   162.4   1.339  0.253
## Residuals   7836 950229   121.3
```

The p-value of 0.25 suggests that we cannot reject the null hypothesis, meaning there is no statistically significant difference in hours worked per week across racial categories for those earning more than $50k. That is, any differences in hours worked between racial groups are likely due to chance, and not due to systematic factors in relation to race.

Q4 Answer: There is no evidence to suggest a statistically significant difference in hours worked per week between racial categories for persons earning more than $50k.