HW 01: Education & median income in US counties

Dav King

2022-09-18

Set up

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(scales)
```

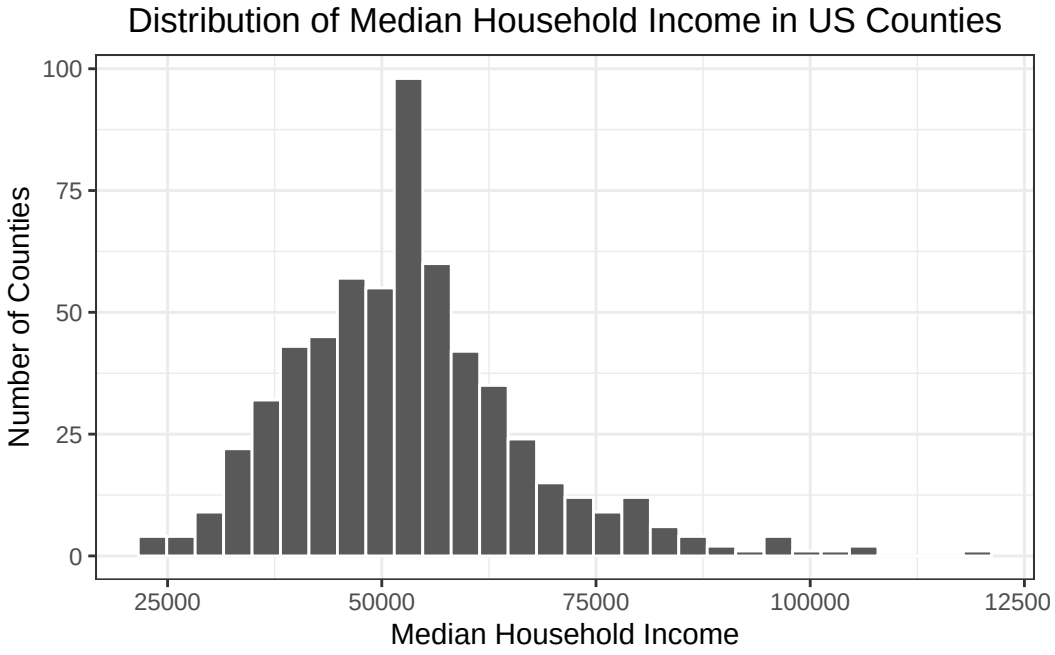
```
county_data_sample <- read_csv("data/us-counties-sample.csv")
map_data_sample <- read_csv("data/county-map-sample.csv")
map_data_all <- read_csv("data/county-map-all.csv")
```

Exercises

Exercise 1

```
ggplot(county_data_sample, aes(x = median_household_income)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(x = "Median Household Income", y = "Number of Counties",
       title = "Distribution of Median Household Income in US Counties") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
fivenum(county_data_sample$median_household_income)
```

```
[1]  23447.0  44387.5  52502.0  59642.5 119730.0
```

The distribution of `median_household_income` is more-or-less normal and unimodal (one could argue that it is right-skewed, but given that it is almost normal, we will consider it

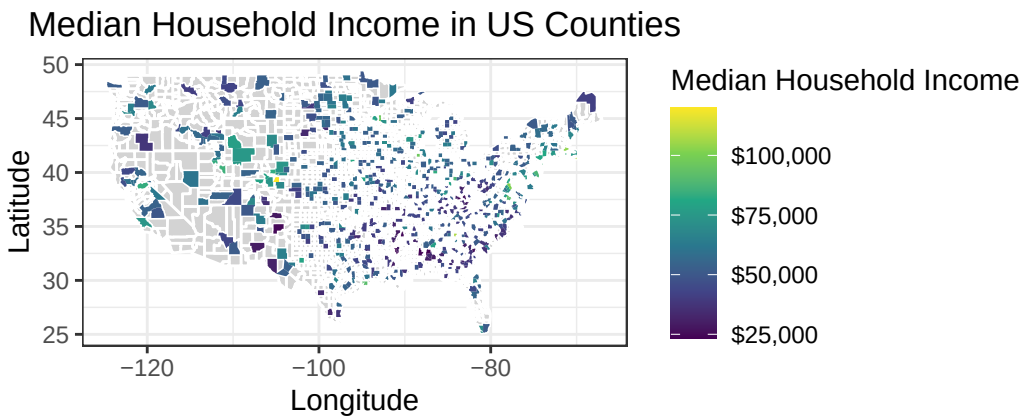
as normal). It is centered at a little over 50000, with a median at 52502, and spread over the range (23447, 119730), with no super obvious outliers (though some of the values close to the maximum, above ~75000, might be outliers).

Exercise 2

```
county_map_data <- left_join(county_data_sample, map_data_sample)
```

```
Joining, by = c("state", "name", "fips")
```

```
ggplot(map_data_all) +
  geom_polygon(aes(x = long, y = lat, group = group),
               fill = "lightgray", color = "white") +
  geom_polygon(data = county_map_data, aes(x = long, y = lat, group = group,
                                           fill = median_household_income)) +
  labs(x = "Longitude", y = "Latitude", fill = "Median Household Income",
       title = "Median Household Income in US Counties") +
  scale_fill_viridis_c(labels = label_dollar()) +
  coord_quickmap() +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



This map shows the distribution of median income in US counties contained within our sample. It shows that most counties have a lower median income, especially across most of the American

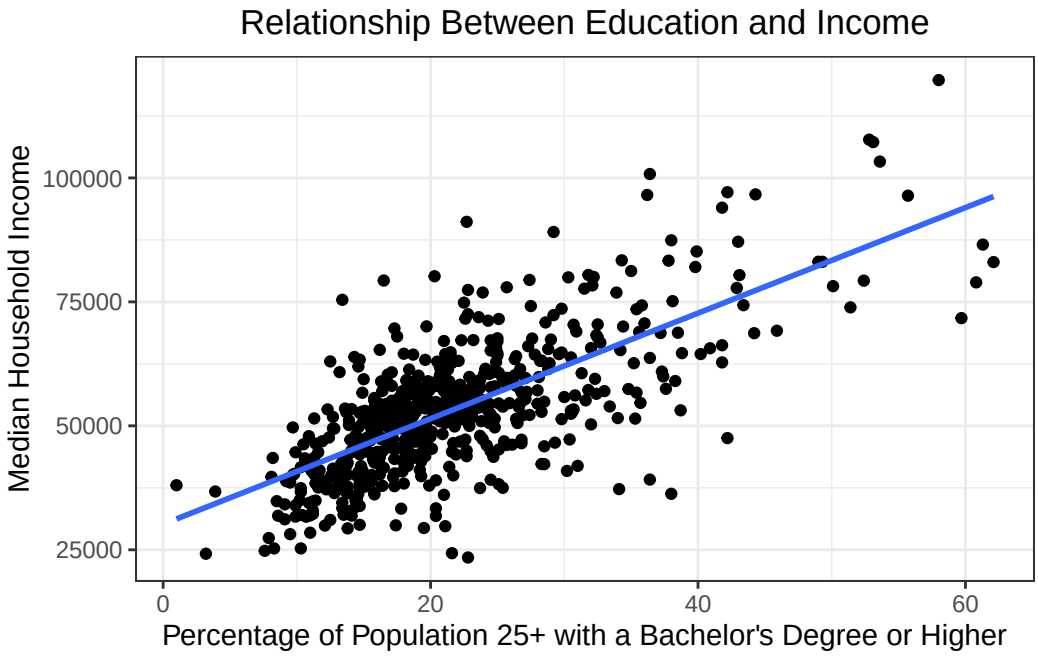
South and Midwest. The conties with higher median incomes seem to be in either the Northeast or the West (focused mostly on Colorado, Wyoming, and Utah).

This map is better at showing a) which counties in the US (and therefore regions in general) are represented in the sample and b) what relationship there might be between urbanism and wealth than the histogram is. However, the histogram is better at showing us the most frequent values in the data, how they are spread, and generally around what set of values the data are centered.

Exercise 3

```
ggplot(county_data_sample, aes(x = bachelors, y = median_household_income)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  theme_bw() +
  labs(x = "Percentage of Population 25+ with a Bachelor's Degree or Higher",
       y = "Median Household Income",
       title = "Relationship Between Education and Income") +
  theme(plot.title = element_text(hjust = 0.5))
```

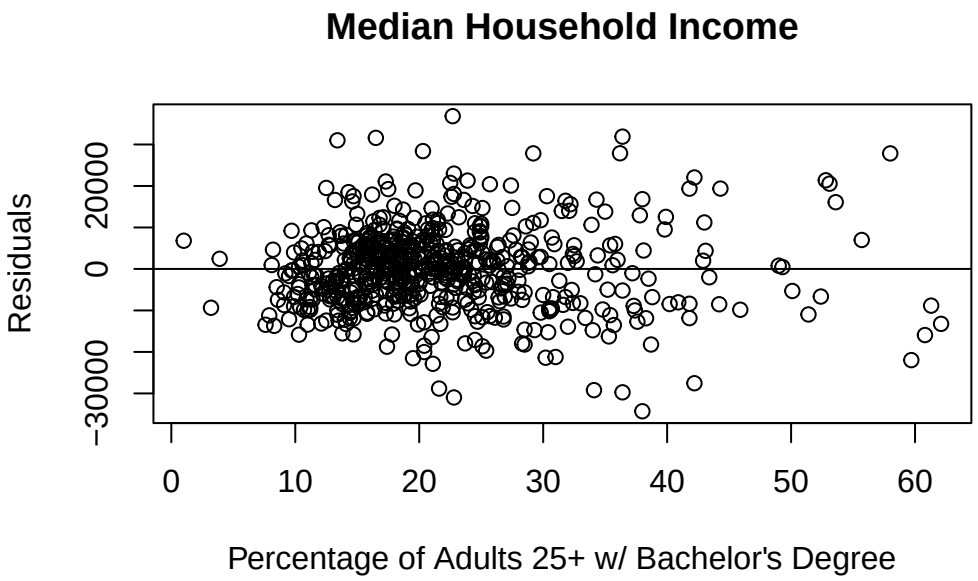
```
`geom_smooth()` using formula 'y ~ x'
```



There generally seems to be a fairly strong positive linear relationship between the two variables. There don't seem to be any major outliers, and the residuals generally seem uncorrelated with the independent variable - however, they do seem to be farther from the line in general at higher values of `bachelors`, which is possibly due to the relative scarcity of data points at such values but is nonetheless worth looking at as we carry out our analysis.

```r
bachModel = lm(median_household_income ~ bachelors, data = county_data_sample)
bachResid = resid(bachModel)

plot(county_data_sample$bachelors, bachResid,
     xlab = "Percentage of Adults 25+ w/ Bachelor's Degree", ylab = "Residuals",
     main = "Median Household Income")
abline(0, 0)
```



Looking at the residual plot, however, there does not seem to be any clear trend in the residuals for this relationship.

Exercise 4

$$median\_household\_income = \beta_0 + \beta_1 \times bachelors$$

Exercise 5

```
incomeModel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(median_household_income ~ bachelors, data = county_data_sample)

incomeModel %>%
  tidy() %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	30155.530	1011.886	29.801	0
bachelors	1064.276	42.879	24.820	0

$$\hat{median\_household\_income} = 30155.530 + 1064.276 \times bachelors$$

Exercise 6

The slope that suggests that, for every one point increase in the percentage of adults aged 25+ in a county who have at least a Bachelor's degree, we would expect to see an increase of $1064.276 in that county's median income. The intercept suggests that, in a county where 0% of adults aged 25+ have earned at least a bachelor's degree, we would expect the median income to be $30155.53. While we wouldn't necessarily expect to see a county with 0% of adults holding a Bachelor's degree, some counties approach this closely enough that the intercept is relevant here.

Exercise 7

The population of interest is the full 3,000+ counties in the US, while the sample is our random subsample of 600 counties. It is reasonable to treat this sample as representative of the population, because it was chosen randomly and thus we should expect it to contain counties with qualities that more or less reflect the true population rates of those qualities.

Exercise 8

$H_0 : \beta_1 = 0$, or the slope of the relationship between the percentage of adults aged 25+ with a bachelor's degree and the median household income in a county is equal to zero.

$H_a : \beta_1 \neq 0$, or the slope described above is not equal to zero.

```
set.seed(8)

nullDist <- county_data_sample %>%
  specify(median_household_income ~ bachelors) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  fit()

obs <- county_data_sample %>%
  specify(median_household_income ~ bachelors) %>%
  fit()

get_p_value(nullDist, obs_stat = obs, direction = "two-sided")
```

```
Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.
```

```
Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.
```

```
# A tibble: 2 x 2
  term      p_value
  <chr>       <dbl>
1 bachelors       0
2 intercept       0
```

The p value of ~0 tells us that we have sufficient evidence to reject H_0 and believe that there is, in fact, a relationship between the percentage of adults aged 25+ who have a bachelor's degree and the median houshold income in a US county.

Exercise 9

```r
set.seed(9)

nullCI <- county_data_sample %>%
  specify(median_household_income ~ bachelors) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  fit()

get_confidence_interval(
  nullCI,
  point_estimate = obs,
  level = .95,
  type = "percentile"
)
```

```
# A tibble: 2 x 3
  term      lower_ci upper_ci
  <chr>        <dbl>    <dbl>
1 bachelors     962.    1173.
2 intercept   27887.   32265.
```

We are 95% confident that the true slope of the relationship between the percentage of adults aged 25+ holding a bachelor's degree and the median household income in a county, that is to say the amount that we expect the median household income to increase by given a one point increase in the percentage of bachelor's degrees, lies on the interval (962, 1173).

Exercise 10

Generally speaking, the hypothesis test and confidence interval do support the consensus - as proven by the hypothesis test, there is a significant relationship between median household income and bachelor's degrees, and the confidence interval failing to contain zero (which we would expect to see if there weren't such a relationship) is further evidence of this fact. What these data tell us is that places that have more adults with bachelor's degrees also make more money, leading us to the conclusion that adults who have completed a bachelor's degree generally earn higher incomes. However, regressing this through county-level data is awkward, unnecessary for that conclusion given that we have more specific data, and so should be taken with a grain of salt when used to support a claim that it itself does not directly test.

Exercise 11

```
compModel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(median_household_income ~ household_has_computer,
      data = county_data_sample)

compModel %>%
  tidy() %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-21174.695	2468.634	-8.577	0
household_has_computer	1065.528	34.966	30.473	0

```
glance(incomeModel)$r.squared
```

```
[1] 0.5074364
```

```
glance(compModel)$r.squared
```

```
[1] 0.6082867
```

```
augBach <- augment(incomeModel$fit)
rmse(augBach, truth = median_household_income, estimate = .fitted)$.estimate
```

```
[1] 9641.769
```

```
augComp <- augment(compModel$fit)
rmse(augComp, truth = median_household_income, estimate = .fitted)$.estimate
```

```
[1] 8598.243
```

These results do in fact support the researcher's claim. Our first look is comparing the standard error of the two models, or the average distance that the observed values fall from the regression line. For the model using computers, this standard error is 34.996, whereas for the model using a bachelor's degree, the standard error is higher at 42.879. This means that the model had bigger residuals or was, on average, off by more for the model using bachelor's degree than the model using computers. Looking at their R^2 values confirms this story - the model using bachelor's degrees has an R^2 of .507, which is a good amount of the variance in median income being explained by the percentage of adults aged 25+ in a county holding a bachelor's degree, but it is ten points lower than the R^2 of .608 held by the model using computer access. Additionally, when comparing the RMSE values of the two models, we again see lower RMSE values for the model using computers - suggesting that the model is off by less, on average, and further defending the researcher's point.