Homework SLR

Arjun Mannem

2023-11-17

### 1a. First, load some R packages for this assignment.

# Load necessary packages  
library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'dplyr' was built under R version 4.3.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

## Warning: package 'flextable' was built under R version 4.3.2

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(corrr)

## Warning: package 'corrr' was built under R version 4.3.2

library(ggfortify)  
library(broom)

### 1b. Set the default ggplot2 and flextable themes using the code below.

# Set ggplot theme for visualizations  
theme\_set(ggthemes::theme\_few())  
  
# Set options for flextables  
set\_flextable\_defaults(na\_str = "NA")  
  
# Load function for printing tables nicely  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")

### 2. Load the state-level ACS data for 2021 into R using the code below.

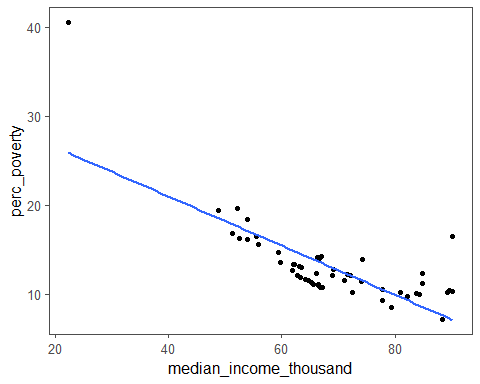
# Importing state-level U.S. ACS data for 2021  
censusData <- read\_csv("https://raw.githubusercontent.com/dilernia/STA418-518/main/Data/census\_data\_2008-2021.csv") %>%   
 dplyr::filter(!stringr::str\_detect(county\_state, pattern = ","), year == 2021) %>%   
 dplyr::mutate(perc\_poverty = 100\*prop\_poverty,  
 median\_income\_thousand = median\_income / 1000)

## Rows: 11421 Columns: 25  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): county\_state  
## dbl (24): year, population, median\_income, median\_monthly\_rent\_cost, median\_...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

### 3. Create a scatter plot including a straight-line of best fit showing the percent of people in poverty (perc\_poverty) by the median income in thousands of dollars for each state (median\_income\_thousand) using the ggplot2 package, reproducing the plot below.

# Creating a scatter plot  
censusData %>%   
 ggplot(aes(x = median\_income\_thousand,   
 y = perc\_poverty)) +   
 geom\_point() +   
 geom\_smooth(method = "lm", se = FALSE)

## `geom\_smooth()` using formula = 'y ~ x'



### 4. Using the lm() function in R, fit a simple linear regression model with median income in thousands of dollars as the predictor variable and percent poverty as the response variable.

# Fitting a simple linear regression model  
slrModel <- lm(perc\_poverty ~ median\_income\_thousand,  
 data = censusData)

### 5. Print the model output using the broom::tidy() and make\_flex() functions to reproduce the output below.

# Printing model output  
slrModel %>%   
 broom::tidy() %>%   
 make\_flex()

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 32.03 | 2.30 | 13.90 | <2e-16 |
| median\_income\_thousand | -0.28 | 0.03 | -8.34 | 5e-11 |

### 6. Provide a statement of the estimated regression equation using median income in thousands of dollars as the predictor variable and percent poverty as the response variable.

A statement of the estimated regression equation is

yHat = 32.03 - 0.28\*x,

where:

yHat is the predicted percent of people in poverty

βHat0 = 32.03 is the estimated intercept

βHat1 = -0.28 is the estimated slope

x is the median income in thousands of dollars for each state

### 7. Provide the value of and interpret the estimated slope in context.

βHat1 = -0.28

*Blueprint Interpretation*

For every 1 unit increase in x, we expect y to increase/decrease by |βHat1|.

*Interpretation in context*

For every 1 thousand dollars increase in the median income of a state, we expect the percent of people in poverty to decrease by 0.28.

### 8. Provide the value of and interpret the estimated intercept in context. Is this appropriate in this context? Why or why not?

βHat0 = 32.03

*Blueprint Interpretation*

The expected value of y given that x=0 is βHat0.

*Interpretation in context*

Here,the expected percent of people in poverty given the median income of a state is 0 dollars is 32.03.

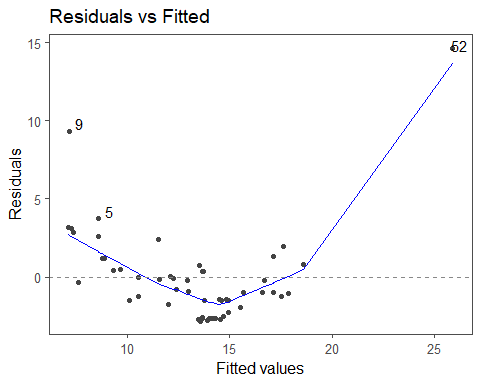
This is inappropriate considering the value of the βHat0.

In this case, if the median income of a state is 0 dollars, this implies more than 50 percent of the people are having an income of 0 dollars.

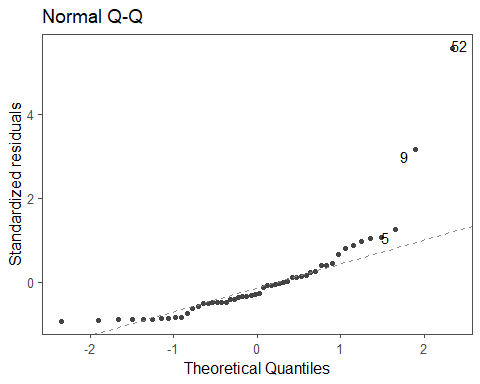
It is also inappropriate since it is considered **extrapolation**, since expected percent of people in poverty given the median income of a state is 0 dollars is beyond the observed data, leading to unreliable predictions..

### 9. Create and display diagnostic plots for the simple linear regression model, including the Residuals vs Fitted plot, the Normal Q-Q plot, Cook’s distance plot, and the Residuals vs Leverage plot.

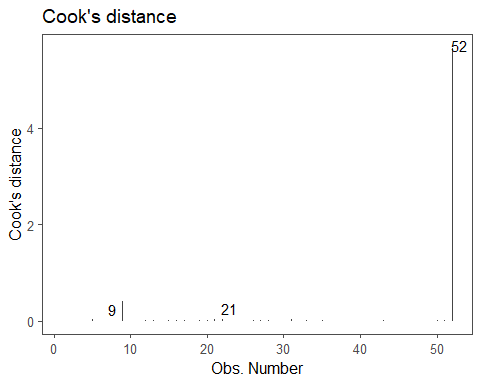
# Creating diagnostic plots  
  
# Residual by fitted value plot  
autoplot(slrModel, which = 1, label.repel = TRUE)@plots[[1]]



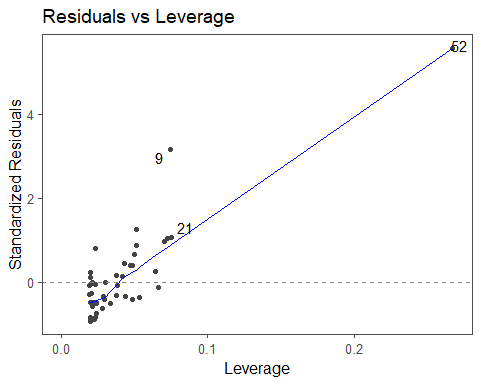
# Quantile-quantile (QQ) plot  
autoplot(slrModel, which = 2, label.repel = TRUE)@plots[[1]]



# Cook's Distance plot  
autoplot(slrModel, which = 4, label.repel = TRUE)@plots[[1]]



# Standardized residuals by leverage  
autoplot(slrModel, which = 5, label.repel = TRUE)@plots[[1]]



### 10. State and check if each assumption for fitting a simple linear regression model is met, providing specific evidence to support your conclusions from the R output obtained.

We assume that we have a random sample.

**Normality:** the residuals must be normally distributed. Checked using a quantile-quantile (QQ) plot of the residuals. Since there are only few points that are away from the line and most of them are on the diagonal line, we can say that the normality assumption is violated.

We will proceed with checking the other assumptions for the sake of the problem.

**Homoskedasticity:** the residuals must have constant variance / spread across all predicted values Checked using the residual by predicted value plot.

There is an outward fanning pattern in the residual by fitted value plot, so the homoskedasticity assumption is violated.

**Independence:** the residuals must be independent of the predicted values. Checked using the residual by predicted value plot.

There is a quadratic pattern in the residual by fitted value plot, so the independence assumption is violated.

**Linearity:** there must be a linear relationship between each predictor and Y, our response variable. Checked using the residual by predicted value plot or a scatter plot between X and Y.

Linearity is met since the points in the scatter-plot follow a linear relationship between the median income of a state in thousands of dollars and the percent of people in poverty.

### 11. Are there any influential points? If so, how many are there? Cite specific evidence to support your conclusion.

Yes, there is one influential point with a Cook’s D value of about ~6(observation 52) which is greater than 0.5. (observation 9 appears to be near to the value of 0.5, it appears lesser when measured using digital ruler.)

### 12. Are there any outliers? If so, how many are there? Cite specific evidence to support your conclusion.

Yes, there are two outliers since they have standardized residuals that are larger than 2 in magnitude(observations 9 and 52).

### 13. Regardless of whether or not assumptions are met, formally state the hypotheses for testing if the regression slope is 0 or not.

Our hypotheses are H0:β1=0 vs. Ha:β1≠0

### 14. Provide the test statistic, p-value, and decision for this hypothesis test.

**test statistic**: t = -8.34

**p-value**: 5e-11

**decision**: Reject the null hypotheses, since the p-value is less than 0.5.

### 15. Interpret the results of the hypothesis test in the context of the problem.

We have sufficient evidence that there is a negative linear relationship between the median income in thousands of dollars for each state and the percent of people in poverty at 5% significance level.