Lab 03

Rohit Kulkarni

1/26/2022

Data: Gift aid at Elmhurst College

In today's lab, we will analyze the `elmhurst` dataset in the openintro package. This dataset contains information about 50 randomly selected students from the 2011 freshmen class at Elmhurst College. The data were originally sampled from a table on all 2011 freshmen at the college that was included in the article "What Students Really Pay to go to College" in *The Chronicle of Higher Education* article.

You can load the data from loading the openintro package, and then running the following command:

```
data(elmhurst)
```

The `elmhurst` dataset contains the following variables:

`family_income`	Family income of the student
`gift_aid`	Gift aid, in ($ thousands)
`price_paid`	Price paid by the student (= tuition - gift_aid)

Exercises

Exploratory Data Analysis

1. Plot a histogram to examine the distribution of `gift_aid`. What is the approximate shape of the distribution? Also note if there are any outliers in the dataset.

The distribution is slightly skewed to the left, with one outlier of gift aid between 30 and 35.

2. To better understand the distribution of `gift_aid`, we would like calculate measures of center and spread of the distribution. Use the `summarise` function to calculate the appropriate measures of center (mean or median) and spread (standard deviation or IQR) based on the shape of the distribution from Exercise 1. Show the code and output, and state the measures of center and spread in your narrative. *Be sure to report your conclusions for this exercise and the remainder of the lab in dollars.*

The function I used for the summary statistics: summary(elmhurst$gift_aid)

results: Min: $7,000 1st Quartile: $16,250 Median: $20,470 3rd Quartile: $23,520 Max: $32,720 IQR: $23,520 - $16,250 = $7270

The median gift aid given by the university is $20,470, with an inter-quartile range of $7,270. There is one one outlier of $32,720.

3. Plot the distribution of `family_income` and calculate the appropriate summary statistics. Describe the distribution of `family_income` (shape, center, and spread, outliers) using the plot and appropriate summary statistics.

The distribution of family income is skewed to the right with a median of $88,060. The IQR is $73,090 with several outliers with much higher income.

4. Create a scatterplot to display the relationship between `gift_aid` (response variable) and `family_income` (predictor variable). Use the scatterplot to describe the relationship between the two variables. Be sure the scatterplot includes informative axis labels and title.

Code used to plot scatterplot:

```
ggplot(data=elmhurst, aes(x=family_income, y=gift_aid) ) + geom_point() +
  xlab('Family Income in thousands') + ylab('Gift Aid in thousands') +
  ggtitle("Gift Aid by Family Income")
```

There is a negative correlation between family income and gift aid. The higher a family's income is, the less gift aid they are likely to recieve.

Simple Linear Regression

5. Use the `lm` function to fit a simple linear regression model using `family_income` to explain variation in `gift_aid`. Complete the code below to assign your model a name, and use the `tidy` and `kable` functions to neatly display the model output. *Replace X and Y with the appropriate variable names.*

```
model <- lm(gift_aid ~ family_income, data = elmhurst)
tidy(model) %>%
  kable(digits = 3)
```

6. Interpret the slope in the context of the problem.

For every $1000 increase in family income, there is a $43 decrease in the gift aid.

7. When we fit a linear regression model, we make assumptions about the underlying relationship between the response and predictor variables. In practice, we can check that the assumptions hold by analyzing the residuals. Over the next few questions, we will examine plots of the residuals to determine if the assumptions are met.

Let's begin by calculating the residuals and adding them to the dataset. Fill in the model name in the code below to add residuals to the original dataset using the `resid()` and `mutate()` functions.

```
elmhurst <- elmhurst %>%
  mutate(resid = residuals(model))
```

8. One of the assumptions for regression is that there is a linear relationship between the predictor and response variables. To check this assumption, we will examine a scatterplot of the residuals versus the predictor variable.

Create a scatterplot with the predictor variable on the x axis and residuals on the y axis. Be sure to include an informative title and properly label the axes.

```
ggplot(data=elmhurst, aes(x=family_income, y=resid) ) + geom_point() +
  xlab('Family Income in thousands') + ylab('Residuals') +
  ggtitle("Residuals of LR on Family Income and Gift Aid")
```

9. Examine the plot from the previous question to assess the linearity condition.

    - *Ideally, there would be no discernible shape in the plot. This is an indication that the linear model adequately describes the relationship between the response and predictor, and all that is left is the random error that can't be accounted for in the model, i.e. other things that affect gift aid besides family income.*
    - *If there is an obvious shape in the plot (e.g. a parabola), this means that the linear model does not adequately describe the relationship between the response and predictor variables.*

Based on this, is the linearity condition is satisfied? Briefly explain your reasoning.

The linearity condition is satisfied because the residual plot has no discernible shape.

10. Recall that when we fit a regression model, we assume for any given value of x, the y values follow the Normal distribution with mean $\beta_0 + \beta_1 x$ and variance σ^2. We will look at two sets of plots to check that this assumption holds.

    We begin by checking the constant variance assumption, i.e that the variance of y is approximately equal for each value of x. To check this, we will use the scatterplot of the residuals versus the predictor variable x. Ideally, as we move from left to right, the spread of the y's will be approximately equal, i.e. there is no "fan" pattern.

    Using the scatterplot from Exercise 8 , is the constant variance assumption satisfied? Briefly explain your reasoning. *Note: You don't need to know the value of σ^2 to answer this question.*

    The constant variance assumption is met because the spread of the residuals is fairly constant at all levels of family income.

11. Next, we will assess with Normality assumption, i.e. that the distribution of the y values is Normal at every value of x. In practice, it is impossible to check the distribution of y at every possible value of x, so we can check whether the assumption is satisfied by looking at the overall distribution of the residuals. The assumption is satisfied if the distribution of residuals is approximately Normal, i.e. unimodal and symmetric.

    Make a histogram of the residuals. Based on the histogram, is the Normality assumption satisfied? Briefly explain your reasoning.

```
hist(elmhurst$resid)
```

The Normality assumption is satisfied because the histogram of residuals appears to be somewhat normal.

12. The final assumption is that the observations are independent, i.e. one observation does not affect another. We can typically make an assessment about this assumption using a description of the data. Do you think the independence assumption is satisfied? Briefly explain your reasoning.

I believe that the independence assumption is satisfied because the gift aid that one family receives should not inherently affect the aid that another family with a different income level should receive.

Using the Model

13. Calculate R^2 for this model and interpret it in the context of the data.

code used to calculate R^2:

```
summary(model)$r.squared
```

The value of R^2 according to the model is 0.2485, which means that 24.85% of the variation in the gift aid can be explained by the family income.

14. Suppose a high school senior is considering Elmhurst College, and she would like to use your regression model to estimate how much gift aid she can expect to receive. Her family income is $90,000. Based on your model, about how much gift aid should she expect to receive? Show the code or calculations you use to get the prediction.

gift_aid = -0.043 * family_income + $24,319 = -0.043 * $90,000 + $24, 319 = $20,449

She should expect to receive $20,449 in gift aid based on the model.

15. Another high school senior is considering Elmhurst College, and her family income is about $310,000. Do you think it would be wise to use your model calculate the predicted gift aid for this student? Briefly explain your reasoning.

It would not be wise to use my model to calculate her predicted aid because her family income is a relative outlier compared to the other family incomes, so the prediction would have a low likelihood of being accurate.

You're done and ready to submit your work! Knit, commit, and push all remaining changes. You can use the commit message "Done with Lab 2!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and that all documents are updated in your repo on GitHub. Then submit the assignment on Gradescope following the instructions below.