STAT 3022: Practical Assignment
Professor Kazeem Adepoju
Kim, Sol

Topic: Relation between the average salary and the sales tax rates in US cities

Introduction

My research topic is the relation between the average salary and the sales tax rates by cities in the US. I wanted to compare the average salary by cities to the income tax rates by cities in the US, at first. I simply expected that if some cities have higher average salaries than other cities, then they might have higher income tax rates than other cities which have lower average salaries. It may sound obvious that the relation between the income tax rates and the average income is a positive correlation, but I wanted to find out more specific information. However, income tax rates were too complicated to expect since income tax rates depend on many variables such as income range, marriage status, etc. For this reason, I changed the topic to searching for a relation between the average salary and the sales tax rates by cities because the sales tax is as important as the income tax. The sales tax affects people's daily spending and living expenses. According to the Tax Foundation, the average taxpayer will pay just over $1,000 per year in sales tax. I assumed that if people get a high-paying job, then they are more likely to spend more money. So the sales tax will be a great revenue to the government, and the government wants to impose higher sales tax because many people can afford that. Therefore, I expected that if some cities have higher average salaries than other cities, they might have more expensive sales tax rates than other cities which have lower average salaries.

I will use a simple regression model for this research to find out the relation between the average income and sales tax rates. The simple linear regression is identifying and fitting a linear model for a quantitative response based on a quantitative predictor (STAT2, 2018). It is a common model to summarize the relationship between two things. To determine whether the relationship between the average monthly net salary and the sales tax rates is statistically significant, I will consider the relationship at a 95% significance level.

Methods and Material

For this research, I gathered average monthly net salary data and sales tax rates by cities in the US. Cities are chosen randomly. The average salary data was collected from the website NUMBEO. It shows the average salary in 2022. The data of sales tax rates was collected from the website Avalara, and the tax rates are based on the 2021 government policy. While I was collecting the data, I realized that the highest monthly net salary is $8895.24 in Bellevue, Washington, and the lowest is $2672.00 in Tucson, Arizona. Bellevue's average monthly net salary is more than 3 times higher than Tucson's. The highest sales tax rate I collected is 10.25%, and the lowest sales tax rate is 0%. There are some cities that have no sales tax. Through some research I found out that some cities are often having alternative financial resources as a sales tax compensation so that they don't have sales tax or impose very small amounts of sales tax . For example, Anchorage in Alaska has no sales tax because Alaska has a lot of oil and gas royalties instead (Dschaak, 2020).

The data provided in SalarySalestax is a random sampling of 87 cities in the US. The first six lines of the data can be seen in Table 1 under the Appendix.

I used the following variables for my research:

1. City: The name of a city in the US which is randomly chosen.
2. State: The name of a state where the following city belongs to.
3. Salary: The average monthly net salary (after tax) by the city.
4. Sales: The sales tax rates of the city.

In order to determine the relation between the monthly income and sales tax, I expected that if citizens have high income, the city might impose more expensive sales tax. And, as vice versa, if citizens have low income, the city might impose lower sales tax. So, the explanatory variable is the average monthly net salary, and the response variable is sales tax rates.

Here are my hypotheses:

- The null hypothesis: there is no relationship between the average monthly salary and the sales tax rates.
- The alternative hypothesis: there is a relationship between the average monthly salary and the sales tax rates.

Results

I made a linear regression model based on my hypothesis and tested it. The Table 2 under the Appendix is coefficients calculated.

According to the coefficients, I can make the fitted regression model as follow :

Sales = 6.1142011 + (0.0003565 * Salary)

This fitted regression model predicts that if the average monthly salary is $1000, then the sales tax rates will be around 6.47% as below :

6.1142011 + (0.0003565 * 1000) = 6.470701

From Table 2, I see that the p-value of this model is 0.01776. Since I am using the 95% confidence level of this research, when I only consider the p-value, I can reject the null hypothesis and say that there is significant evidence that the average monthly salary and the sales tax rates are related to each other. However, the R-squared value is shown as 0.06509. That is, only 6% of the variance in the Sales variable can be explained by the Salary variable. This is too low to confirm that the two variables are associated with. Therefore, to check whether this simple linear model is reasonable and trustable, I did a few more tests. A key tool for fitting a model is to compare the values it predicts for the individual data cases to the actual values of the response variable in the dataset. The differences in predicting each

response can be measured through the residual (STAT2, 2018). Here are some results of testing residuals of a fitted model :

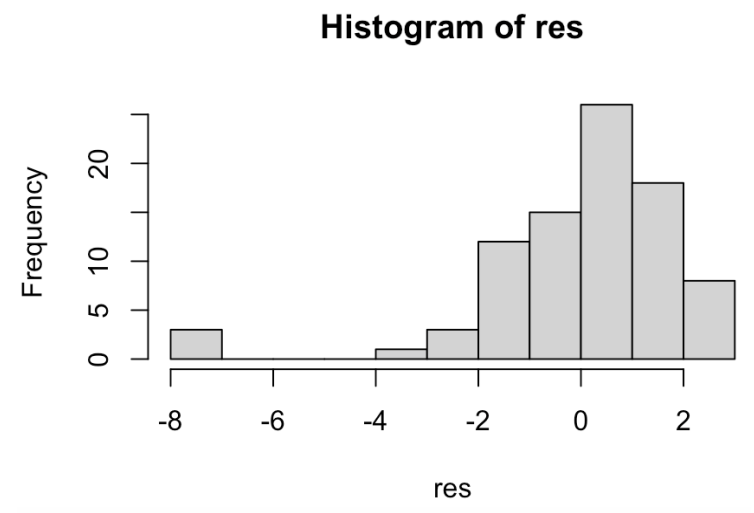

Table 3. Histogram of residuals of the fitted model

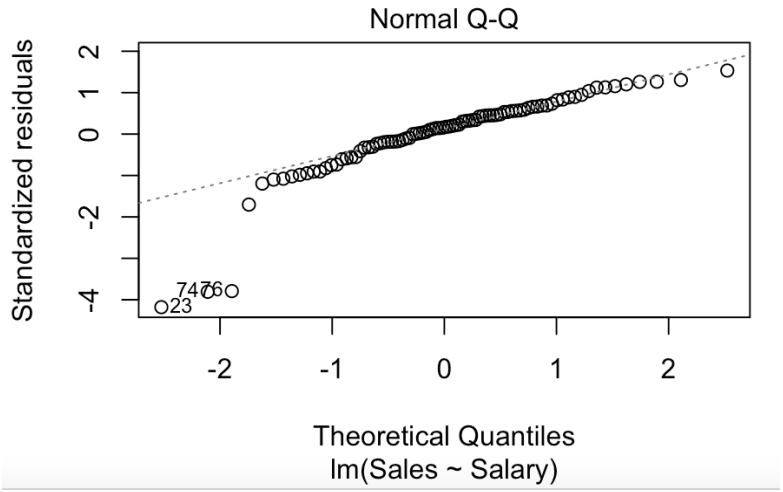

Table 4. QQnorm and QQline of residuals

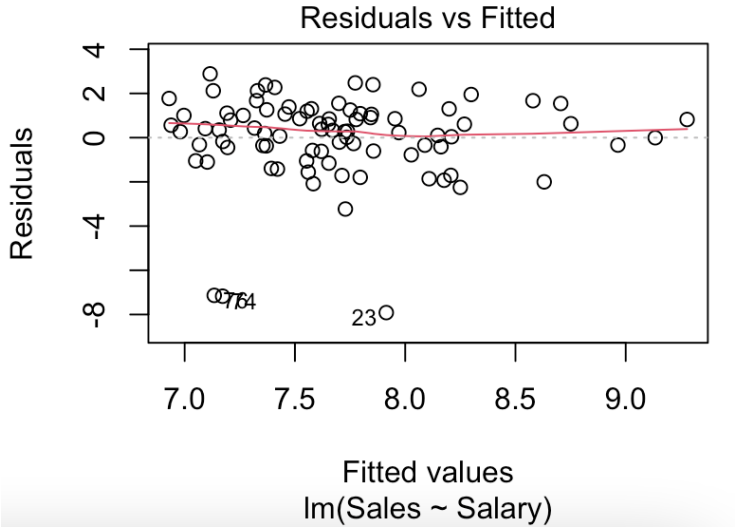

Table 5. Residuals vs Fitted values

Through the results in Table 3, Table 4, and Table 5, I see that the residuals are left skewed and there are some outliers in values. A skewed residuals distribution would imply that my model is biased. The outliers can disturb statistical analysis, so if the outliers are due to known errors, they should be removed from the data before a more detailed analysis is performed (TIBCO Spotfire, n.d.). The predictable reason for errors is that in the dataset, there are few cities that have no sales tax for each reason. Thus, I slightly changed my data by excluding cities which have no sales tax from my dataset. Through this change, I get new coefficients and residual results like Table 6 under Appendix, Table 7, Table 8, and Table 9.

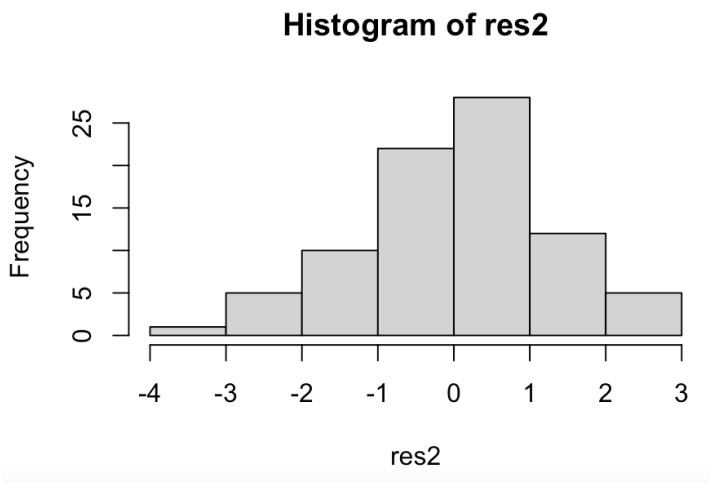

Table 7. Histogram of adjusted residuals

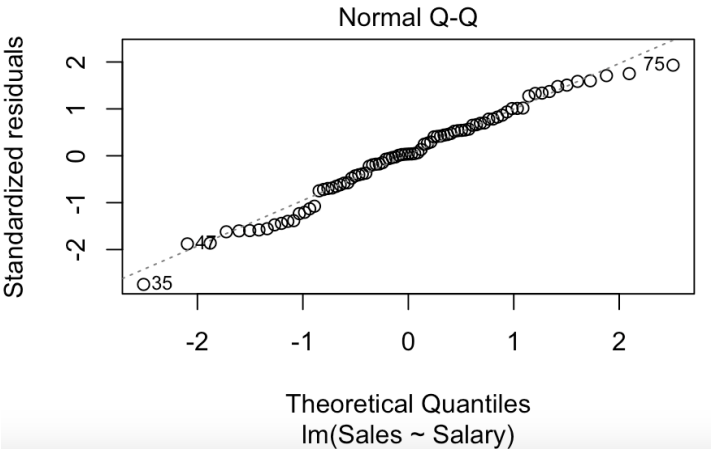

Table 8. QQnorm and QQline of adjusted residuals

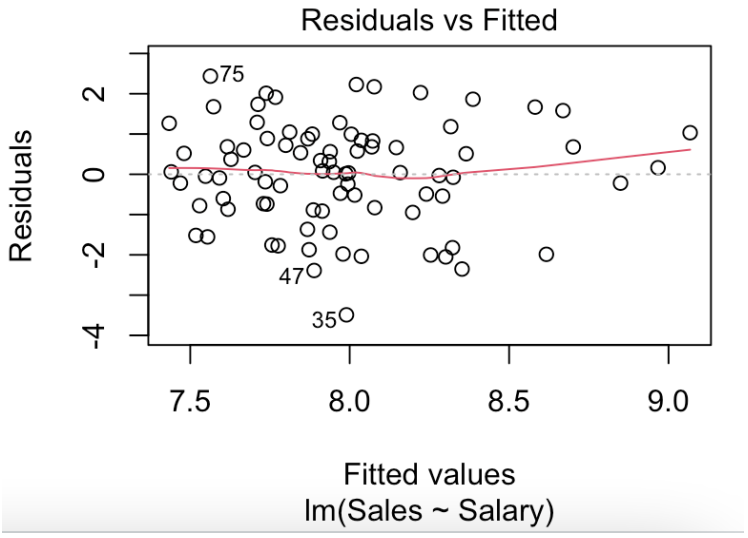

Table 9. Residuals vs Fitted values

The changed model looks better than before, but there are still some outliers in residuals. However, since I can't find a reason right now, I will skip it. Based on the adjusted coefficients, I can make new regression model as follow :

Sales = 6.8668076 + (0.0002479 * Salary)

When I tested this model, the p-value was 0.01571, so I can't reject the null hypothesis. To check the normality of the dataset, I did the Shapiro-Wilk test and got 0.2369 as the p-value like Table 10 under the Appendix. Since it's greater than the level of significance, 0.05, I see that the variables are not normally distributed. Since the normality is violated and the data is skewed, I tried to do data transformation by taking the log. The log transformation reduces or removes the skewness of the original data, improves linearity between variables, and boosts validity of the statistical analyses (Htoon, 2020). The response variable, the sales tax rates, is log-transformed, and the fitted model will be like follow regarding to the Table 11 under Appendix :

log(Sales) = 1.931e+00 + (2.931e-05 * Salary)

I can interpret this model as, every one unit increase in Salary is multiplied by about 1.000029, and for every one unit increase in Salary, Sales increases by about 0.0029%. The p-value from the Shapiro-Wilk normality test from Table 12 under Appendix is 0.008468, so I can say the variables are now normally distributed. However, the intercept and the slope of the model seem rarely effective. The R-squared value is 0.05711 which means that 5.7% of the variance in the Sales variable can be explained by the Salary variable. This is not a significant number of explained variance, so I can confirm again that the relationship between the two variables are weak. Lastly, the p-value for this model is 0.05532 which is slightly greater than the level of significance. Consequently, I can't reject the null hypothesis, and in

other words, I can say there is no relationship between the average monthly net salary and the sales tax rates.

Discussion/Conclusion

My research topic was started from my curiosity about the relation between people's income and the tax rates. I specified the research topic to the relation between the average monthly net salary and the sales tax rates by cities in the US. My hypothesis was that if the city has a higher average monthly income than the other cities, then they will impose higher sales tax rates compared to the other cities. When I chose the research topic and planned which statistical method to use, I only thought about the simple linear regression method because it is easy to do and perfect for analyzing the relation between two variables. However, while I collected the data and analyzed it, I found some unexpected errors such as zero sales tax rates for some cities, normality error, and so on. To fix the issues, I tried to make some changes such as excluding the known outliers, log transformation, etc. Through these changes, I could get a better result than before.

As a result of my research, there was no significant evidence that the average monthly net salary and the sales tax rates are related. Although this result contradicts my hypothesis, I predicted that the monthly income and the sales tax rates are not related while I was collecting data. I believed the relationship between the average income and the sales tax rates was definitely a positive correlation before this research because when I traveled to rich neighborhoods and poor neighborhoods, I always noticed the price difference between two. I learned that there is no significant relationship between the average income and the sales tax rates and also learned about some special cities and states where they don't impose the sales tax for some reason through this research.

References

Cannon, A.R., Cobb, G.W., Hartlaub, B.A., Legler, J.M., Lock, R.H., Moore, T.L., Rossman, A.J., & Witmer J.A. (2018). *STAT2: modeling with regression and ANOVA Second Edition.* W. H. Freeman.

NUMBEO. (n.d.). *Average Monthly Net Salary (After Tax) (Salaries And Financing) by City.* <https://www.numbeo.com/cost-of-living/prices_by_city.jsp?displayCurrency=USD&itemId=105>

Dschaak, Casey. (2020, October 29.) *Policy Brief: A History of Alaska Oil Taxes and How They Work.* <https://alaskapolicyforum.org/2020/10/history-alaska-oil-taxes/>

TIBCO Spotfire. (n. d.) *Available Diagnostic Visualizations.* <https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/prd/prd_available_diagnostic_visualizations.htm>

Htoon, Kyaw Saw. (2020, February 29). *Log Transformation: Purpose and Interpretation.* <https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9>

Appendix

```
> head(SalarySalestax)
# A tibble: 6 × 4
  City          State Salary Sales
  <chr>         <chr>  <dbl> <dbl>
1 Bellevue      WA     8877. 10.1
2 Santa Clara   CA     8471.  9.13
3 San Francisco CA     7996.  8.63
4 San Jose      CA     7398.  9.38
5 Fremont       CA     7269. 10.2
6 Jersey City   NJ     7058.  6.63
```

Table1. Head of the dataset

```
Call:
lm(formula = Sales ~ Salary)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9142 -0.5996  0.3104  1.0791  2.8847

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.1142011  0.6858050   8.915 8.78e-14 ***
Salary      0.0003565  0.0001474   2.418   0.0178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.907 on 84 degrees of freedom
Multiple R-squared:  0.06509,   Adjusted R-squared:  0.05396
F-statistic: 5.848 on 1 and 84 DF,  p-value: 0.01776
```

Table 2. Test result of the Simple linear regression model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.8668076  0.4703654  14.599   <2e-16 ***
Salary      0.0002479  0.0001005   2.468   0.0157 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6. Adjusted coefficients

```
        Shapiro-Wilk normality test

data:  res
W = 0.98041, p-value = 0.2369
```

Table 10. Result of the Shapiro-Wilk test

```
Call:
lm(formula = log(Sales) ~ Salary)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55973 -0.09656  0.01860  0.11699  0.28921

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.931e+00  6.195e-02  31.174   <2e-16 ***
Salary      2.931e-05  1.323e-05   2.215   0.0296 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1686 on 81 degrees of freedom
Multiple R-squared:  0.05711,   Adjusted R-squared:  0.04547
F-statistic: 4.906 on 1 and 81 DF,  p-value: 0.02957
```

Table 11. Test result of the log transformation model

```
        Shapiro-Wilk normality test

data:  res
W = 0.98041, p-value = 0.2369
```

Table 12. Result of the Shapiro-Wilk test