Homework 2

Jaehyun (Alex) Lee

Regression and Multivariate Data Analysis (STAT-UB.0017)

Prof. Jeffrey Simonof

We commonly relate or even joke about frequent consumption of fast food as being of somewhat of a low socioeconomic class. Because fast food is so cheap and accessible, I thought that people with demands of low-priced food would be more inclined to consume fast food as opposed to people who are more indifferent about the price of food. Since people of lower socioeconomic levels would seem to have higher demand for cheaper food compared to people of relatively higher socioeconomic levels, I became curious if there is a significant relationship between the demand of fast food and socioeconomic levels in corresponding regions.

To quantify my questions, I have set up my dependent and independent variables as something that intuitively represent demand for fast food and socioeconomic levels in corresponding regions. My dependent variable is the number of McDonald’s per 100,000 people in 50 different states, and independent variable is GDP (in millions of dollars) per 100,000 people in 50 different states. The data on number of McDonald’s per 100,000 people in different states were given from the source below but in order to match the data of GDP per capita in different states to number of McDonald’s per 100,000 people in different states, I multiplied 100,000 to the average GDP per capita for every state. For each of the 50 states in the sample, the GDP (in millions) per 100,000 people is used as a potential predictor of the number of McDonald’s per 100,000 people (Number of McDonald’s proportional to the population).

(I have attained data about McDonald’s from the following website: <http://247wallst.com/consumer-products/2016/09/04/the-number-of-mcdonalds-in-all-50-states/>. And the data of GDP per capita from the following website: <https://en.wikipedia.org/wiki/List_of_U.S._states_by_GDP_per_capita>. I used both data as of 2015.)

The relationship I want to examine can be represented using a simple regression model which is,

where number of McDonald’s per 100,000 people in a state = (No. of McDonald’s in that state) / (state’s total population / 100,000), and GDP per 100,000 people in a state = GDP per capita in a state x 100,000 (represented in millions of dollars).



Looking at the scatterplot, there is no noticeably strong linear relationship but there seems to be a rough negative association between the two variables. Also, there is one state towards the top of the scatterplot which has noticeably high number of McDonald’s relative to its GDP, which is Ohio. Doing a least squares regression with McDonald’s per 100,000 people in a state as the dependent variable, and GDP (in millions) per 100,000 people in a state as independent variable provides the following analysis:

Regression Analysis: MD per 100k ppl in a state versus GDP (in millions of $) per 100k ppl

Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
| Regression | 1 | 6.142 | 6.1422 | 10.96 | 0.002 |
| GDP(in millions $) for 100k ppl | 1 | 6.142 | 6.1422 | 10.96 | 0.002 |
| Error | 48 | 26.895 | 0.5603 |  |  |
| Total | 49 | 33.037 |  |  |  |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 0.748535 | 18.59% | 16.90% | 12.87% |

Coefficients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | Coef | SE Coef | T-Value | P-Value | VIF |
| Constant | 6.995 | 0.580 | 12.07 | 0.000 |  |
| GDP(in millions $) for 100k ppl | -0.000391 | 0.000118 | -3.31 | 0.002 | 1.00 |

Regression Equation

|  |  |  |
| --- | --- | --- |
| MD per 100k ppl | = | 6.995 - 0.000391 GDP (in millions $) for 100k ppl |

Fits and Diagnostics for Unusual Observations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs | MD per 100,000 ppl | Fit | Resid | Std Resid |  |
| 35 | 7.100 | 5.165 | 1.935 | 2.61 | R |

*R  Large residual*

The regression is not as strong as I have expected with R-sq of 18.59%, indicating that predictor is accounting for approximately 19% of the variability in the number of McDonald’s in the corresponding state proportional to its population. However, p value of 0.002 in the F-test seems to imply strong evidence against the null hypothesis, indicating strong degree of significance of the regression. The result is a little bit confusing because it doesn’t sound intuitive for a regression to have low R-sq value while also having a low p value which indicates the regression is statistically significant (x variable has some predictive power to the y) but at the same time x variable does not account for majority of variability in y. The combination of low R-sq and low p value makes me question if the significant predictor is still meaningful or not, but I will explore further what that could mean in the following paragraph.

Note that the intercept in this case is not meaningful as 0 input in the independent variable would mean that about 7 McDonald’s would exist for 100,000 people in a certain state despite the fact that people do not have money to purchase food from the store. The slope coefficient is negative but it is a value that is very close to 0, implying that a million dollar increase in the GDP among 100,000 people in a state is associated with a very small negative change (0.000391 number of stores) in the number of McDonald’s for 100,000 people in that respective state. It is intuitive that the slope coefficient is so tiny because a million dollar change in the GDP of 100,000 people is proportionately small compared to collective GDP of 100,00 people and thus it would merely have any effect on the number of McDonald’s around those people. However, the point that should be emphasized is the negative relationship between the predictor and target variable as these statistically significant variables (since p value is close to zero) confirm the existence of this negative relationship although the association between the two variables is not noticeably strong. Thus it could be said that although the slope coefficient is very tiny and close to 0, the p value of either F test and t test that is close 0 enables us to reject the null hypothesis and accept that the slope coefficient is in fact significant and not zero. Although the value of the coefficient is close to zero, it is still significant and the negative association between the two variables exist on a statistically significant level.



The standard error of the estimate is 0.75, telling us that this model could be used to predict the number of McDonald’s for every 100,000 people in a certain state to within + or – 1.5 stores roughly 95% of the time. The 95% confidence interval given above provides answer to the question of what is the best estimate for the true average number of McDonald’s for 100,000 people in every state that have GDP for 100,000 people in the corresponding state of some value? The definition of the confidence interval indicates that 95% of the intervals calculated on my sampling of the data in 50 states would capture the true population mean.

Prediction for MD per 100k ppl

Regression Equation

|  |  |  |
| --- | --- | --- |
| MD per 100k ppl | = | 6.995 - 0.000391 GDP(in millions $) per 100k ppl |

Settings

|  |  |
| --- | --- |
| Variable | Setting |
| GDP(in millions $) per 100k ppl | 5500 |

Prediction

|  |  |  |  |
| --- | --- | --- | --- |
| Fit | SE Fit | 95% CI | 95% PI |
| 4.84501 | 0.132346 | (4.57891, 5.11111) | (3.31663, 6.37338) |

To quantitatively see what the CI and PI mean, I have chosen a specific value of my predictor of 5500 million dollars for a group of 100,000 people to see what the corresponding number of McDonald’s around those people would be. The values of confidence interval, (4.579, 5.111), indicates the range where the average number of McDonald’s per 100,000 people would lie for all the states with collective GDP per 100,000 people of 5500 million dollars. Prediction interval on the other hand provides an estimate of what the number of McDonald’s per 100,000 people would be 95% of the time for one state with GDP per 100,000 people of 5500 million dollars. As you can see, 95% PI is displaying a range of + or – 2s, which in this case equals 1.5 as mentioned above.

There is one point that stands out towards the top which lies beyond the prediction interval and that is Ohio. That particular point is unusual and in order to check if it is truly unusual and to determine if the assumptions of the least squares regression hold, we have to examine the following plots:



The assumptions seem to hold for the most part:

1. In the plot of the residuals vs the fitted values, there is no apparent pattern to it but there is a point at the top that stands out and could be unusual; therefore it would be good to examine that point.
2. Plots or residuals vs observation likewise does not display any pattern
3. The data does not have time structure to it thus it is not necessary to plot it against time
4. In the normal plot of the residuals, points are roughly along the straight line but again, there is one point at the top right (Ohio) that stands out and seems a bit unusual.

Although Ohio seems as if it is quite noticeably high in the plot of residual vs fitted value, it is not in fact an outlier. I have run an outlier test to determine if that point was an outlier as shown in the following analysis:

Outlier Test: MD per 100k ppl, GDP(in millions $) per 100k ppl

Method

|  |  |
| --- | --- |
| Null hypothesis | All data values come from the same normal population |
| Alternative hypothesis | Smallest or largest data value is an outlier |
| Significance level | α = 0.05 |

Grubbs' Test

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | Mean | StDev | Min | Max | G | P |
| MD per 100k ppl | 50 | 5.108 | 0.821 | 3.500 | 7.100 | 2.43 | 0.635 |
| GDP(in millions $) per 100k ppl | 50 | 4827 | 906 | 3163 | 6771 | 2.15 | 1.000 |

**\* NOTE \* No outlier at the 5% level of significance**

I was worried that if Ohio were to be an outlier, it would have a strong influence over the fitted slope and intercept, giving a poor fit to the bulk of data points. Although Ohio is not an outlier, it is still worth noticing that Ohio has relatively many McDonald’s given its comparatively high socioeconomic level. But this raises a question that confuses me a little: is it possible for a point outside of 95% prediction interval (in which the range is determined by plus or minus 2 standard errors) not be an outlier? At this stage, I do not have the intuition to answer the question, and this question made me doubt if Ohio truly does not have a noticeable effect on the regression. Thus, I ran the regression analysis without Ohio to check if there were any noticeable changes:

Regression Analysis: MD per 100k ppl\_1 versus GDP(in ... ) per 100k p\_1

Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
| Regression | 1 | 5.918 | 5.9177 | 12.06 | 0.001 |
| GDP(in millions $) per 100k p\_1 | 1 | 5.918 | 5.9177 | 12.06 | 0.001 |
| Error | 47 | 23.070 | 0.4909 |  |  |
| Total | 48 | 28.988 |  |  |  |

Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 0.700609 | 20.41% | 18.72% | 14.47% |

Coefficients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | Coef | SE Coef | T-Value | P-Value | VIF |
| Constant | 6.921 | 0.543 | 12.74 | 0.000 |  |
| GDP(in millions $) per 100k p\_1 | -0.000384 | 0.000111 | -3.47 | 0.001 | 1.00 |

Regression Equation

|  |  |  |
| --- | --- | --- |
| MD per 100k ppl\_1 | = | 6.921 - 0.000384 GDP(in millions $) per 100k p\_1 |

Fits and Diagnostics for Unusual Observations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs | MD per 100k ppl\_1 | Fit | Resid | Std Resid |  |
| 12 | 4.200 | 5.560 | -1.360 | -2.00 | R |
| 20 | 6.300 | 4.814 | 1.486 | 2.15 | R |
| 43 | 3.800 | 5.225 | -1.425 | -2.06 | R |

*R  Large residual*

The result above only excludes Ohio, the observed unusual point, and as a result, the analysis have altered slightly. The most noticeable changes are the changes in p value and R-sq. As a result of taking out Ohio, p value has decreased from 0.002 to 0.001 while the R-sq has gone up from 18.59% to 20.41%. These aren’t big changes, and changes in a direction that implies a stronger regression relationship. Beta 1 and 0 have changed (slope coefficient from -0.000391 to -0.000384 and constant from 6.995 to 6.921) as well but quite marginally, not affecting the predictive model at a significant level. Another thing to note is that from taking out Ohio, unusual observations have changed as well. Before, the only unusual observation was Ohio, however after it was taken out, the unusual observation became 3 other points (Utah, Maryland, and Idaho). This implies that previously, Ohio’s extremity may have drawn out the regression analysis such that other potential unusual observations were shadowed. In a sense, a “tighter” regression regards previously not an unusual observation as being unusual.



In the residual plots however the unusual points do not stand out as much as Ohio did (in fact they almost look usual). And I cannot deduce, through simple regression, what caused Ohio to have relatively many McDonald’s location despite its comparatively high socioeconomic level. Also, since my regression is based on my theory that number of McDonald’s in different states may be inversely correlated with the socioeconomic levels of corresponding region, there is no studies that I can refer to in order to obtain any information of why the unusual observations are the way they are. The information that is available however is that Ohio was ranked 28th in GDP per capita among 50 states in 2015. It may have been the popularity or geographical characteristics that enabled relatively higher number of McDonald’s to open for its socioeconomic level compared to other states, but I cannot say what factor is associated with this phenomenon for sure. From the regression analysis, we can say that the model displays a statistically significant inverse relationship between GDP of 100,000 people (in millions) of a state and number of McDonald’s per 100,000 people in the corresponding state, but cannot claim that this model is a complete predictive model. As we have seen, Ohio was outside of our prediction interval, and although after taking Ohio out gave us a model that was slightly more statistically significant, there still were some unusual observations. Thus we see that the statistical model I built to examine my theory that socioeconomic levels in different regions should be inversely associated with number of McDonald’s in the same region is not a complete predictive model; however, from the regression we could confirm the existence of statistically significant inverse relationship between the variables and realize that socioeconomic level is not a predictor that can provide single-handed explanation

n to the number of McDonald’s existing in that certain state.