Mark Nemat

STA 3064

Written Report for Project A

1. **Motivation:** My objective for Project A is to determine if cities in San Mateo have a significant amount of citizens who use the Transit to commute to work. This topic is important to me because I use the FSU bus system and ride my bike to get around campus/Tallahassee. A Regression analysis would beneficial, because I can examine the influence of one independent variable on a dependent variable.
2. **Data Description:** I got my data from the County of San Mateo Performance website. This data from this site describes the commute to Work Data by Mode of Transportation for Walk, Bike, Transit. It was calculated as part of the Metropolitan Transportation Commission‘s Bay Area Transportation Survey, 2000. The only data I used from this website is their recorded Transit Numbers and the Jurisdiction. My response variable is the Transit Number and my predictor variable is the population in each Jurisdiction. I found this data from Google the population in each Jurisdiction. My response and predictor variable are quantitative but the Jurisdictions are qualitative. There will be 20 observations in my data table.

<https://performance.smcgov.org/Livable-Community/Commute-to-Work-Data-by-Mode-of-Transportation-for/8xpx-w48e/data>

1. **Data Exploration:**

Data Transit;

Length Jurisdiction $19;

Input @1 Jurisdiction $19. @21 Population 6. @20 Transit 4.;

datalines;

Woodside 5510 7

South San Francisco 67733 2680

San Mateo 105025 2931

San Carlos 30364 557

San Bruno 43047 1656

Redwood City 86200 1899

Portola Valley 4598 0

Pacifica 38759 1684

Millbrae 22557 502

Menlo Park 34549 614

Hillsborough 11444 112

Half Monn Bay 12973 96

Foster City 34151 568

East Palo Alto 29519 688

Daly City 107008 8858

Colma 1504 54

Burlingame 30467 1157

Brisbane 4693 117

Belmont 27113 527

Atherton 7187 100

;

run;

proc print data=transit;

run;

proc sgscatter data=transit;

compare Y=transit X=Population/loess; /\*loess uses local weighted regression to fit a smooth curve through points in a scatter plot.\*/

run;

/\*This code will allow me to print out the slope an intercept

in my plot with the regression line\*/

proc reg data=Transit;

model transit = population;

ods output ParameterEstimates=PE;

data \_null\_;

set PE;

if \_n\_ = 1 then call symput('Int', put(estimate, BEST6.));

else call symput('Slope', put(estimate, BEST6.));

run;

proc sgplot data=transit noautolegend;

title "Regression Line with Slope and Intercept";

reg y=transit x=population;

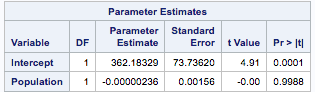
inset "Intercept = &Int" "Slope = &Slope" /

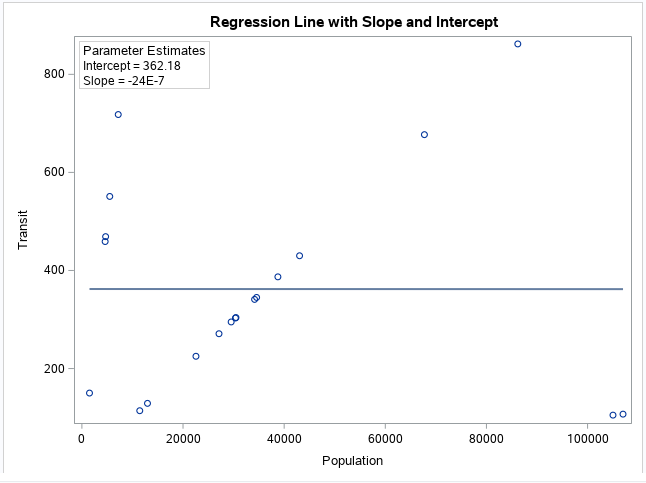
border title="Parameter Estimates" position=topleft;

run;

Y = -.00000236 + 362.18329

R-Squared = 0. With R-Squared and the slope = 0, this indicates that population does not significantly predict transit.





**\*note:** By looking at my residual plots, you can see how my data points are grouped into three populations. From the X vs. Y plot with the regression line, there is a linear pattern when population is approximately < 10,000, 10,000 < population < 100,000, and population > 100,000. I messaged Professor Ramsier about this and he suggested that I group one of the populations and evaluate that separate regression. With that being said, My regressions for Problem 4 will be answered from the data with population around 10,000 < population < 100,000.

Data Transit;

Input Population Transit;

datalines;

67733 2680

30364 557

43047 1656

86200 1899

38759 1684

22557 502

34549 614

11444 112

12973 96

34151 568

29519 688

30467 1157

27113 527

;

run;

proc sgscatter data=transit;

compare Y=transit X=Population/loess; /\*loess uses local weighted regression to fit a smooth curve through points in a scatter plot.\*/

run;

/\*This code will allow me to print out the slope an intercept

in my plot with the regression line\*/

proc reg data=Transit;

model transit = population;

ods output ParameterEstimates=PE;

data \_null\_;

set PE;

if \_n\_ = 1 then call symput('Int', put(estimate, BEST6.));

else call symput('Slope', put(estimate, BEST6.));

run;

proc sgplot data=transit noautolegend;

title "Regression Line with Slope and Intercept";

reg y=transit x=population;

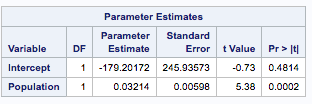
inset "Intercept = &Int" "Slope = &Slope" /

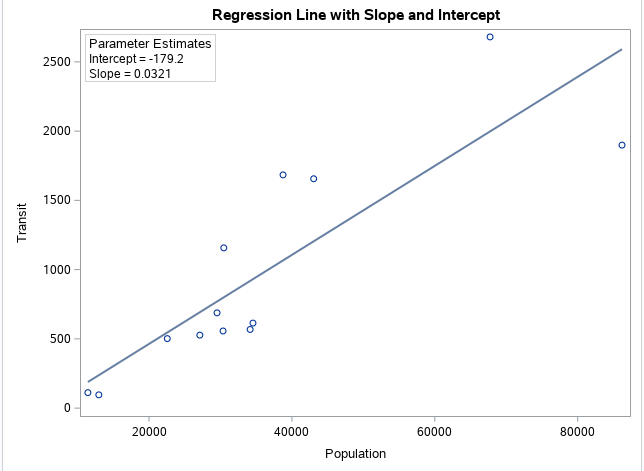
border title="Parameter Estimates" position=topleft;

run;

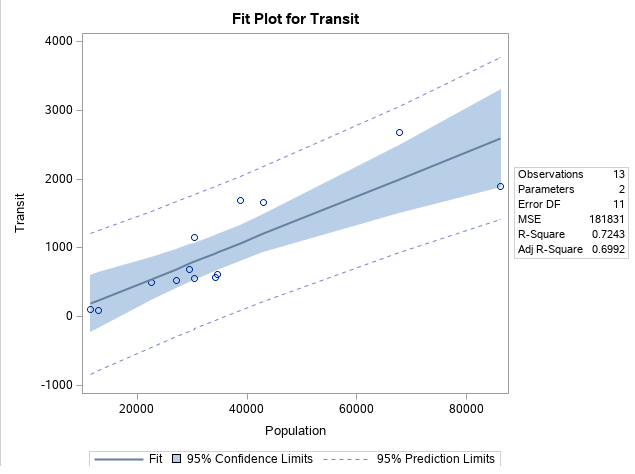
1. **Model Fitting and Analysis:**
2. Fit a simple linear regression model. Provide the model equation.

Y = .03214x -179.20172





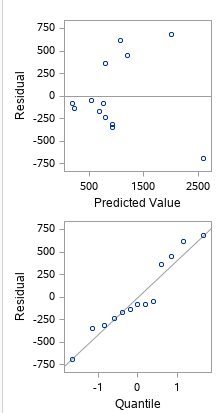
1. Interpret the R2 for your fit.



R-Squared Value = .7243

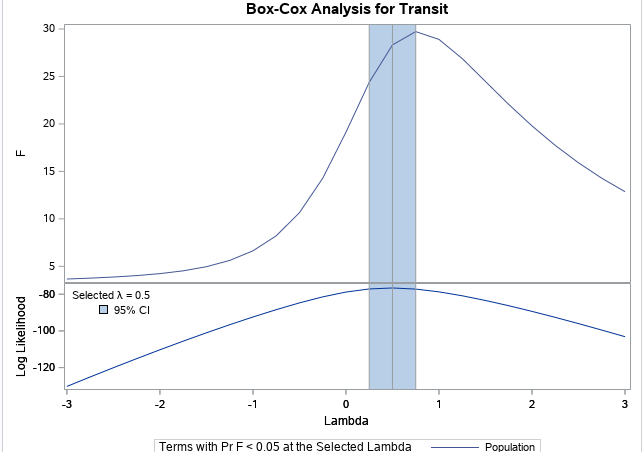
This means that 72.42% of the variance of transit can be predate from the variable population.

c. Perform a residual analysis to determine if all model assumptions are being met.

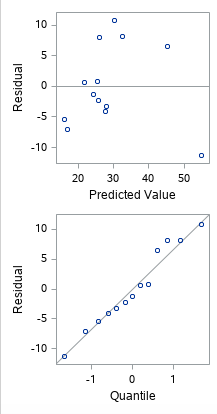


By looking at the Residual vs PV plot, It looks like non-constant variance in your residual vs. predicted plot (megaphone-shaped) as the spread seems to get bigger going left to right. The Linearity is met in the first plot because the data points are scattered around the plot. By looking at the Residual vs Quartile plot, I would say that the normality is met. There is a little bit of a curve to the point that could interfere with the normality. I noticed that were are a few outliers.

1. Explore potential transformations on your response variable (even if your residual  analysis indicates a transformation is not necessary just to confirm). If a transformation is indicated, transform your response, refit your model, and assess.

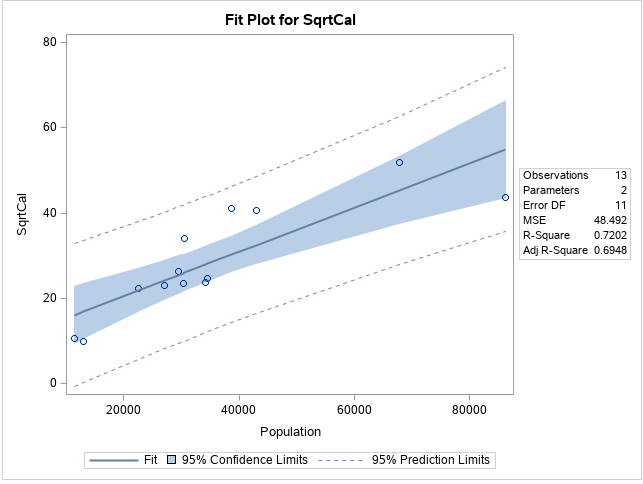


When looking at the Box-Cox plot, the lambda is = .5, which suggests that there would be a transformation. If lambda = 1, then it would be fine. After trying some transformations on Y, such as, LogCal, SqrtCal, InvCal, and ISRCal, as well as, applying it to the regression model; I found that the SqrtCal is the best fit.



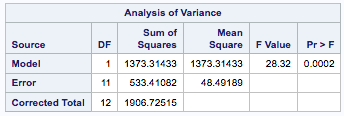
By looking at the Residual vs PV plot, It still looks like non-constant variance in your residual vs. predicted plot as the spread seems to get bigger going left to right. The graph shows a megaphone-shape. The Linearity is met in the first plot because the data points are scattered around the plot. The Residual vs Quartile plot, I would say that the normality is met. This plot looks better after the transformation I noticed that were are a few outliers.

e. For your simple regression model produce a 95% confidence interval for the true slope of your regression line. Interpret.



By looking at the 95% confidence interval on the regression, the data looks fairly linear and most are in the 95% confidence limit and a few points just outside of it but still in-between the 95% prediction limits. This shows a 95% confidence interval for the coefficient.  This is very useful as it helps understand how high and how low the actual population value of the parameter might be.  The confidence intervals are related to the p-values such that the coefficient will not be statistically significant if the confidence interval includes 0.

f. Conduct the ANOVA F-test for the slope. Discuss whether this test indicates that the simple linear model is effective.

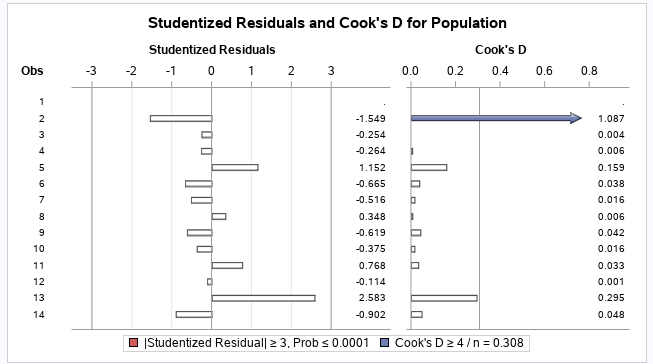


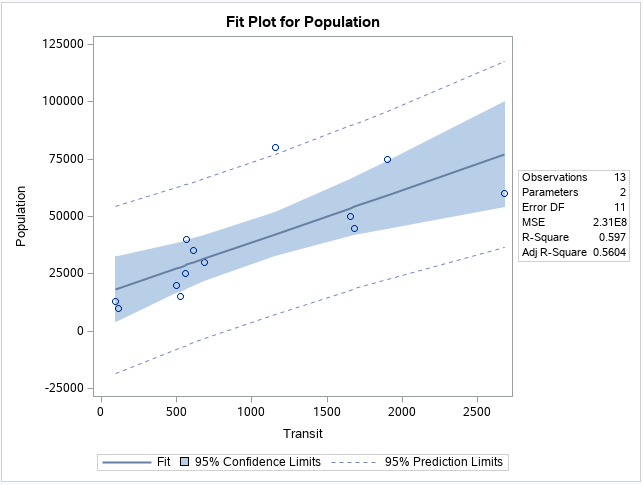
Test statistic = 28.32

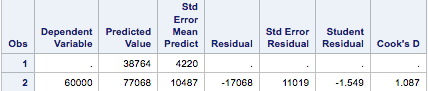
P-Value = .0002 < .05 so reject null hypothesis.

Since p-value is less than the significance level, my data provide sufficient evidence to conclude that the regression model fits the data better and the linear model is effective.

**g.** Select a new value (*x\**) of interest for your predictor variable (preferably a value not contained in your original data set). Produce both 95% confidence and prediction intervals around the predicted response for *x\**. Interpret both intervals.







By looking at the plot, it seems to be linear, however, there is a outlier as you can see a point outside the 95% prediction limits.

The predicted response is 38764.

1. **Conclusion:**

The objective of this project was to determine if cities in San Mateo have a significant amount of citizens who use the Transit to commute to work. My initial assumption was that the larger city population will have more transit usage. By gathering my data and putting it through a regression model, I was able to understand the influence city population had on people who use the transit. When we looked at my original data set, I discovered that population size doesn’t not significantly predict transit. When I grouped my data to cities with 10,000 – 100,000 population size, I found that it does have an impact on the transit, however, I wouldn’t say it’s significant.