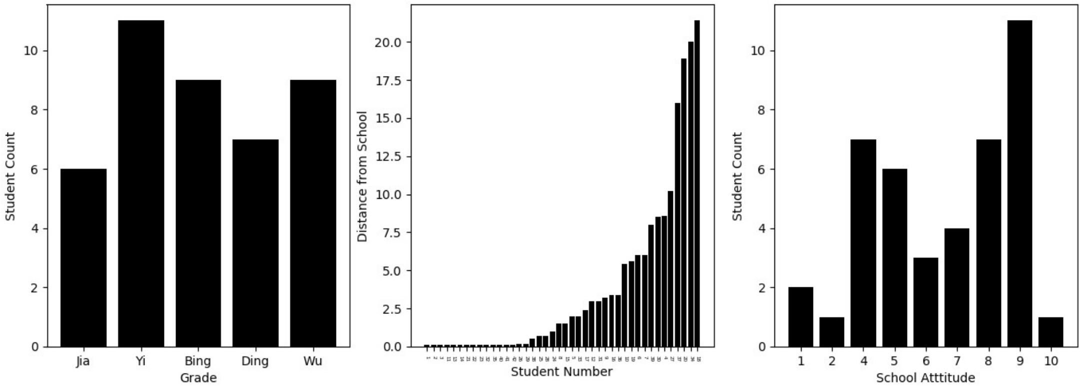
Morley J Weston | Homework 3: Discrete Choices

Question 1: School performance
To explore this data, some initial analysis was done in Python. I also used this to convert all values from Unicode to ASCII to use in R later on. All code can be found on Github.[1]

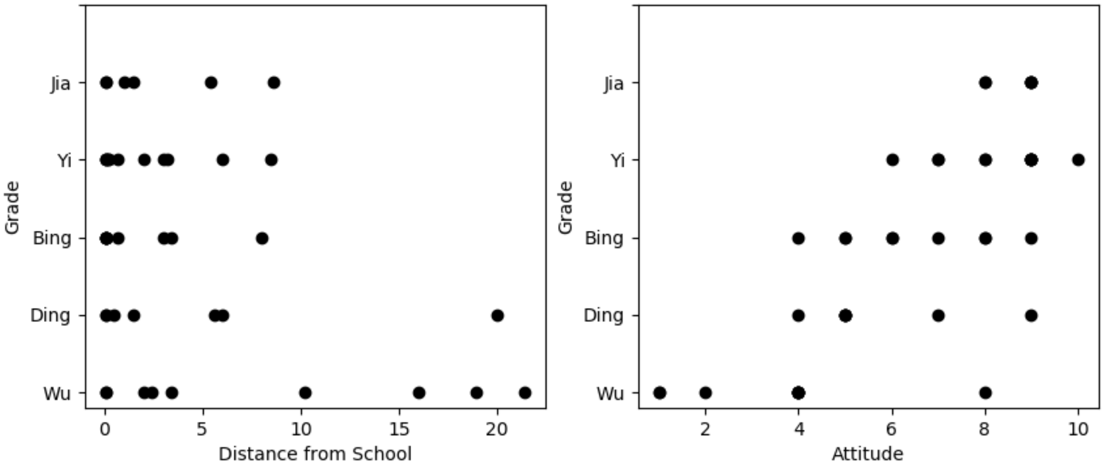
For this question, we are given a list of 42 students' scores, categorized as 甲乙丙丁 or 戊. The other variables given are sex, family income, distance from school, whether they pay tuition, and their attitude toward studying. A first look at the data reveals that the data follows reasonably normal distribution patterns:



Additionally, there are 23 boys and 19 girls in class, 25 pay tuition and 17 do not, and that 21 students come from low-income households, 12 from middle-income, and 9 from high.

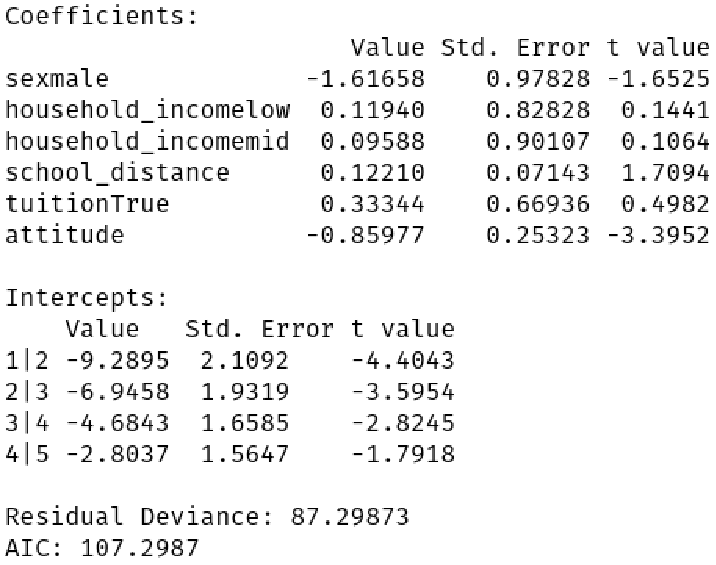
Of these variables, sex and tuition payment can be considered unordered variables, class attitude, household income and grade can be considered ordered logit variables, and distance from school could be considered a continuous variable.

Looking at each variable separately, we can assign each 甲乙丙丁戊 a grade rank of 1-5, with 1 being the best. We can find that boys have a mean grade rank of of 2.7 and girls 4.11, which is a major difference between the variables. Low-income students score 3.38, middle-income students 2.42, and high-income students 2.89, which is a moderate difference. Tuition paying and nonpaying students are almost exactly the same, with score ranks of 3.04 and 3.06, respectively. We can take a look at the more granularized data with the following graphs:



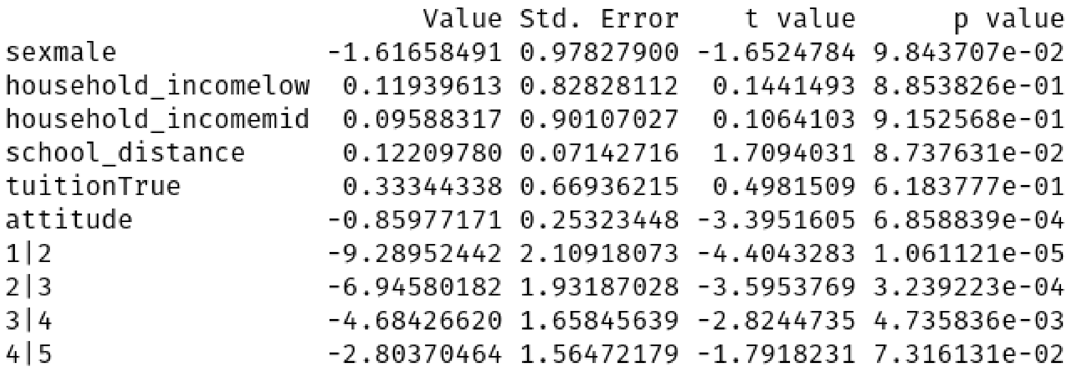
[1] https://github.com/fgorkhs/transportgeography_hw3

Regression:
This data explorations is only intended to give a feel for the data and to check for any anomalies, but a proper regression analysis can show what factors are really affecting students' grades.[2] Using R's *polr* ordinal logistic regression model, we get the following results (Grades are coded 甲=1,乙=2 etc.):



```
Coefficients:
                       Value Std. Error t value
sexmale              -1.61658    0.97828 -1.6525
household_incomelow   0.11940    0.82828  0.1441
household_incomemid   0.09588    0.90107  0.1064
school_distance       0.12210    0.07143  1.7094
tuitionTrue           0.33344    0.66936  0.4982
attitude             -0.85977    0.25323 -3.3952

Intercepts:
    Value   Std. Error t value
1|2 -9.2895  2.1092    -4.4043
2|3 -6.9458  1.9319    -3.5954
3|4 -4.6843  1.6585    -2.8245
4|5 -2.8037  1.5647    -1.7918

Residual Deviance: 87.29873
AIC: 107.2987
```

The T values are very high for sex, distance from school, and attitude, which means that they are less likely to be significant factors. However, without a significance test, it is hard to say for sure. P-values can be calculated as:



```
                          Value Std. Error    t value      p value
sexmale             -1.61658491 0.97827900 -1.6524784 9.843707e-02
household_incomelow  0.11939613 0.82828112  0.1441493 8.853826e-01
household_incomemid  0.09588317 0.90107027  0.1064103 9.152568e-01
school_distance      0.12209780 0.07142716  1.7094031 8.737631e-02
tuitionTrue          0.33344338 0.66936215  0.4981509 6.183777e-01
attitude            -0.85977171 0.25323448 -3.3951605 6.858839e-04
1|2                 -9.28952442 2.10918073 -4.4043283 1.061121e-05
2|3                 -6.94580182 1.93187028 -3.5953769 3.239223e-04
3|4                 -4.68426620 1.65845639 -2.8244735 4.735836e-03
4|5                 -2.80370464 1.56472179 -1.7918231 7.316131e-02
```

As suspected, attitude, distance from school, and sex have statistically significant effects on grade performance, with very low P values. Tuition payment and income levels, however, are do not strongly affect school performance in this study. With the odds ratio, we can assume that an increase of the attitude score of 1 would predict a grade increase of 0.42, for example.

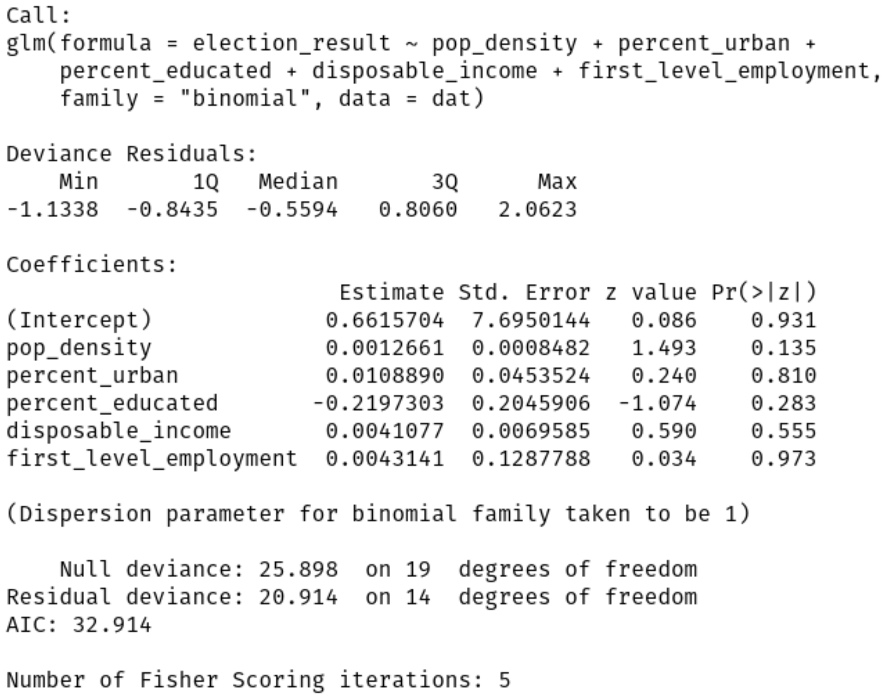
These results make intuitive sense. Students who are farther from school have less time in the day to do schoolwork, and are likely to be disconnected from their classmates. Household income means that a student would have fewer opportunities for extra study programs,

"Attitude" is in a sense a proxy for what a teacher feels about a student; students who are well-behaved are more likely to get a higher attitude score, as well as do better on the tests. Of course, we don't know how objective these tests are, or how much influence a teacher might have over them.

[2] Much of the R code was adapted from: https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

Question 2: Election Results

This is a much smaller data set, analyzing some demographic data and its effect on electoral results. I analyzed this again using R's *polr* package using a binary logit regression, which produced the following results:



```
Call:
glm(formula = election_result ~ pop_density + percent_urban +
    percent_educated + disposable_income + first_level_employment,
    family = "binomial", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1338  -0.8435  -0.5594   0.8060   2.0623

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.6615704  7.6950144   0.086    0.931
pop_density             0.0012661  0.0008482   1.493    0.135
percent_urban           0.0108890  0.0453524   0.240    0.810
percent_educated       -0.2197303  0.2045906  -1.074    0.283
disposable_income       0.0041077  0.0069585   0.590    0.555
first_level_employment  0.0043141  0.1287788   0.034    0.973

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.914  on 14  degrees of freedom
AIC: 32.914

Number of Fisher Scoring iterations: 5
```

From this, we can deduce the following:

- Every additional person per square kilometer increases the log odds of voting non-blue by 0.001.
- Every NT$1000 of disposable income likewise increases the log odds of voting non-blue by 0.04.

The other values did not seem to pass the threshold of statistical significance, but there may be other factors at play.

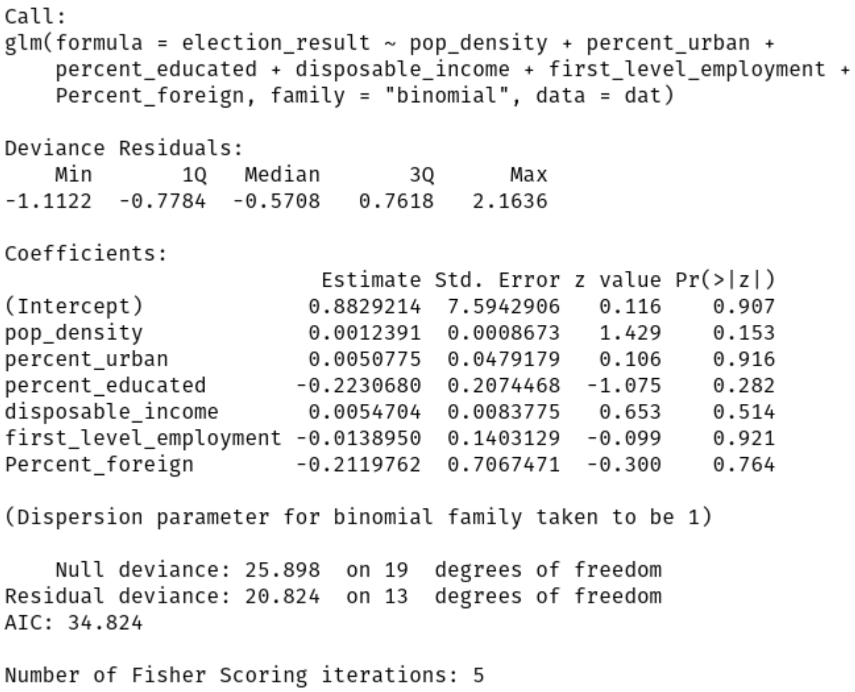
Part 2:

For the second part of this question, I tested the electoral results against the percentage of foreigners in a given county or city, which was obtained through the National Immigration Agency[3] for foreign population and the Ministry of the Interior[4] for the total population. This data was quite surprising; Hsinchu County turned out to be the most international place in Taiwan, followed closely by Taoyuan.

The data for foreigners was processed using the same binary logit regression, producing the results detailed opposite:

[3] https://www.immigration.gov.tw/5475/5478/141478/141380/206756/

[4] https://www.ris.gov.tw/app/portal/346



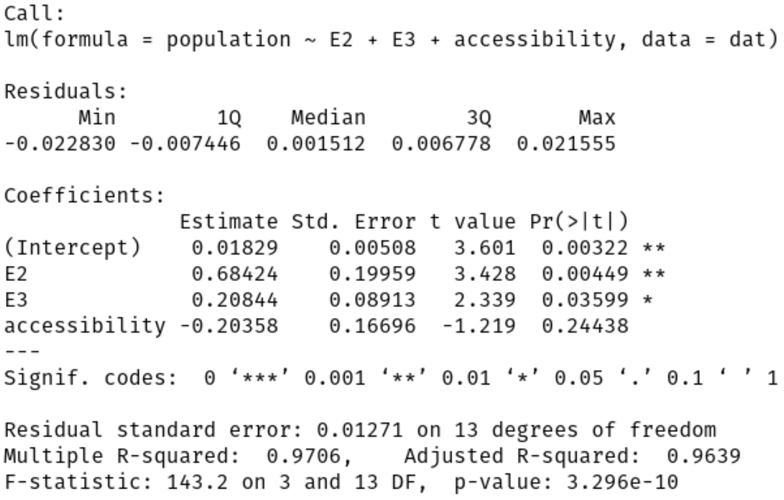
```
Call:
glm(formula = election_result ~ pop_density + percent_urban + 
    percent_educated + disposable_income + first_level_employment + 
    Percent_foreign, family = "binomial", data = dat)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.1122  -0.7784  -0.5708   0.7618   2.1636  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.8829214  7.5942906   0.116    0.907
pop_density            0.0012391  0.0008673   1.429    0.153
percent_urban          0.0050775  0.0479179   0.106    0.916
percent_educated      -0.2230680  0.2074468  -1.075    0.282
disposable_income      0.0054704  0.0083775   0.653    0.514
first_level_employment -0.0138950  0.1403129  -0.099    0.921
Percent_foreign       -0.2119762  0.7067471  -0.300    0.764

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.824  on 13  degrees of freedom
AIC: 34.824

Number of Fisher Scoring iterations: 5
```

As we can see, the percentage of foreigners has no statistically significant effect on voting outcomes. In retrospect, this makes sense; the foreign population in Taiwan is a diverse group, split between students and professionals in the big cities, factory workers in the industrial belt, and a smattering of English teachers and spouses around the country. I initially suspected that foreign population would be much more concentrated in a few places, and might have interesting correlations with voting outcomes, but this was not in fact the case.

Question 3: Accessibility

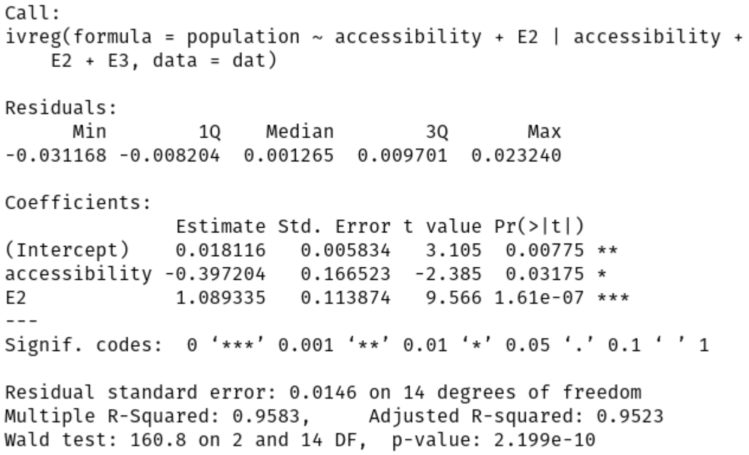
In order to determine the relationship between population, accessibility, secondary industry and tertiary industry, I will use two-stage least squares analysis. In this scenario, secondary industry and accessibility are exogenous variables, in which accessibility only affects population, and secondary industry only affects tertiary industry. Population and tertiary industry affect each other, and are therefore endogenous variables.

To start, a simple Ordinary Least Squares linear regression is performed on the data, which produces some statistically significant results, but they could be stronger:



```
Call:
lm(formula = population ~ E2 + E3 + accessibility, data = dat)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.022830 -0.007446  0.001512  0.006778  0.021555 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)    0.01829    0.00508   3.601  0.00322 **
E2             0.68424    0.19959   3.428  0.00449 **
E3             0.20844    0.08913   2.339  0.03599 * 
accessibility -0.20358    0.16696  -1.219  0.24438   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01271 on 13 degrees of freedom
Multiple R-squared:  0.9706,	Adjusted R-squared:  0.9639 
F-statistic: 143.2 on 3 and 13 DF,  p-value: 3.296e-10
```

In order to perform a two-stage least squares analysis, I used the Advanced Econometrics in R (AER) package. This produced more statistically significant results, demonstrating a strong link between secondary industry and population, but filtered through tertiary industry.



```
Call:
ivreg(formula = population ~ accessibility + E2 | accessibility +
    E2 + E3, data = dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.031168 -0.008204  0.001265  0.009701  0.023240

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.018116   0.005834   3.105  0.00775 **
accessibility -0.397204   0.166523  -2.385  0.03175 *
E2             1.089335   0.113874   9.566 1.61e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0146 on 14 degrees of freedom
Multiple R-Squared: 0.9583,     Adjusted R-squared: 0.9523
Wald test: 160.8 on 2 and 14 DF,  p-value: 2.199e-10
```

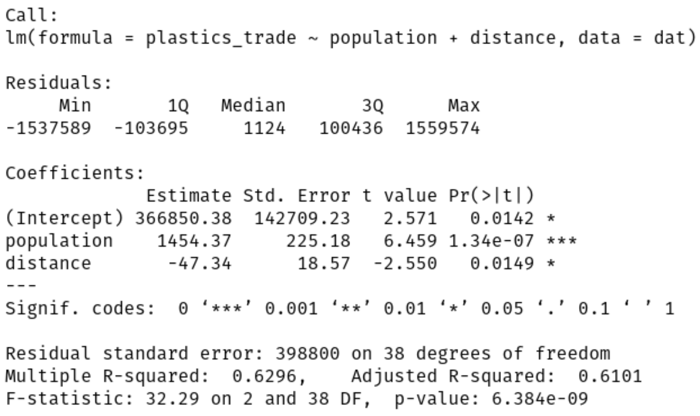
In this scenario, we can deduce that secondary industry gives rise to tertiary industry, which then attracts more people. When this is accounted for, accessibility also becomes a slightly statistically significant factor.

Question 4: Exports

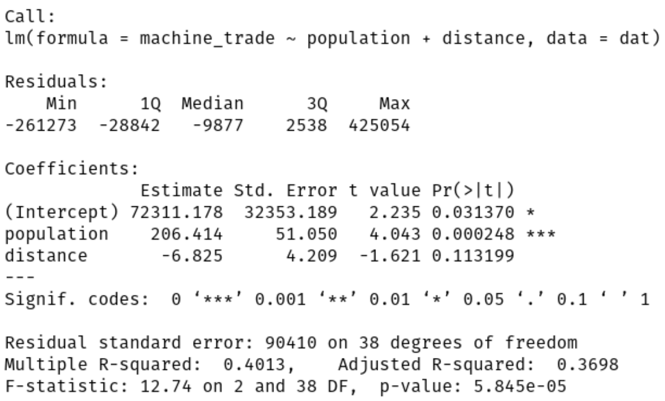
In this final scenario, we will use the gravity model of international trade to show how much a country's population and distance affect it's attractiveness as a trading partner.

We were also asked to provide our own estimates of shipping distances and population for each country listed. For the distance estimates, I used sea-distances.org[5], which calculates the approximate shipping distances between countries. Switzerland, a landlocked country, was not in this database, so shipping distance to Italy and train distance to Zurich was substituted. Population estimates were taken from the World Bank.[6]

Two goods were modeled in this scenario; plastics and electrical machinery. For each of these goods, k, α and β are represented as the estimated intercept, population, and distance coefficients, respectively.



```
Call:
lm(formula = plastics_trade ~ population + distance, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1537589  -103695     1124   100436  1559574

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 366850.38  142709.23   2.571   0.0142 *
population    1454.37     225.18   6.459 1.34e-07 ***
distance       -47.34      18.57  -2.550   0.0149 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 398800 on 38 degrees of freedom
Multiple R-squared:  0.6296,    Adjusted R-squared:  0.6101
F-statistic: 32.29 on 2 and 38 DF,  p-value: 6.384e-09
```



```
Call:
lm(formula = machine_trade ~ population + distance, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-261273  -28842   -9877    2538  425054

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72311.178  32353.189   2.235 0.031370 *
population    206.414     51.050   4.043 0.000248 ***
distance       -6.825      4.209  -1.621 0.113199
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90410 on 38 degrees of freedom
Multiple R-squared:  0.4013,    Adjusted R-squared:  0.3698
F-statistic: 12.74 on 2 and 38 DF,  p-value: 5.845e-05
```

From these linear regressions, we can see that population has a much greater effect on the trade of plastic goods; an increase in 1 million people in a country corresponds with 1454.37 megatons more plastics sold, compared to 206 more megatons of machinery.

[5] https://sea-distances.org/

[6] https://data.worldbank.org/indicator/SP.POP.TOTL

We can also see that the sale of plastics are sensitive to distance, whereas the sale of machinery is not, based on statistical significance. This makes intuitive sense; plastic goods are produced in a variety of countries, so consumers are more likely to buy goods closer to their own countries. Machinery, on the other hand, is a specialized good that is less sensitive to distance, with many consumers around the world buying specifically from Taiwan.