Lab2_w203: Transit Costs Projects

Team Eels - Parastoo Javadi, Stephen Bridwell, Prakash Krishnan, Justin Wong

Last Updated 12/06/2021 10am PST

Contents

1 Introduction

Investing in rapid transit systems is very popular across the globe and such systems oftentimes drive economic value by driving urban growth. Rapid transit systems include subways, metros, and light rails. In addition, a rapid transit system in dense urban population centers can provide low pollution transportation options as building more and more roads is not feasible.

Unlike roadway projects that are relatively small, light rail projects have many more distinct parts including complex project management, design, civil structures, tracks, signaling systems, maintenance facilities, etc. However, despite these similarities, the cost per kilometer is very different across continents.

In this paper, we aim to identify the key contributing features that impact the cost of a rapid transit project. Given that the cost per kilometer of a rapid transit system in New York, USA is 20 times more than Seoul, South Korea, we are particularly interested in exploring the different factors that affect the cost of a rapid transit project and how they differ across geographic locations.

The target audience for this study will include:

- Direct audience: Researchers, planners, journalists, advocates, elected officials, and others interested in contextualizing transit-infrastructure costs and fighting for better projects.

The remainder of this paper is organized in the following way. Section 2 states our key research questions. Section 3 describes the data background and origination, how we developed a model framework using the available data, and the variables influencing the models we developed. Section 4 summarizes the research design and the entire process of the experiments done in the design. Section 5 briefly performs exploratory data analysis for additional context regarding the infrastructure projects dataset, motivating how the model was build and the regression analysis. Section 6 tests the classical linear model assumptions for the best Global, Asia, and Europe model from the previous section. Section 7 discusses the different models created and resulting comparisons across the chosen models. Section 8 elaborates on certain statistical and structural limitations of the models. Finally, in Section 9, we discuss the main key findings and future implications of our results.

2 Research Question

The intent of our research project is to investigate:

1. What are the key contributing features that impact the cost of a rapid transit project?
2. How do factors affecting the cost of a rapid transit project differ across different geographic locations?

3 Data and Variable Definitions

3.1 Data Background

To investigate the research question, we studied the work done by researchers at NYU Marron Institute.

The dataset for the analysis is from https://transitcosts.com/data/ and maintained by Transit Costs Project. The dataset includes data collected from over 600 transit projects across the globe and includes 58 countries and totals more than 11,000 km of urban rail built since the late 1990s.

For each infrastructure project, data points included the location of the project, cost of a project, length of the transit line in kilometers, design features such as a number of stations, tunnels, elevated tracks, and length of project time from start to finish. The costs were adjusted for purchasing power parity (PPP), inflation, and currency exchange. This allows for the comparison of costs across transit projects and also across the globe.

In reviewing the background literature (https://transitcosts.com/cases/), we developed the following model framework to answer the research question.

3.2 Model Framework from Data

We recognize from the literature review that type of project management, politics, resource level, and skills play an important role in project execution and management and ultimately to project duration and costs. Meanwhile, some variables in the dataset appear to be colinear and proxies for one another, representing the same or similar attribute of a project.

3.2.1 Included in our Model Framework

In order to simplify model complexity for this research study, the following factors are considered to understand the influence of the project cost:

- Location of the Project
- Length of the Project in Terms of Kms of line length
- Duration for the project
- Tunnel Length in the Project as tunnel increases costs significantly
- Number of Stations in the Project

3.2.2 Excluded from our Model Framework

Additionally, the typical project cost of a rapid-transit project includes both tangible and intangible costs. Because such features are often difficult to measure, such costs were not available in the dataset and not factored in the modeling. They include:

- Project Management
- Design and Engineering
- Delivery
- Politics
- Resource level

3.3 Variable Definitions

Project years is a calculated field representing the difference between the start year and end year of each project as an integer.

Tunnel length is the length of the tunnels on the project measured in km and represented as a decimal.

Stations is the number of stations on the project and represented as an integer.

Overall length is the length in service, excluding non-revenue tracks toward railyards and train tracks measured in km and represented as a decimal.

Realcost includes all construction and construction-related expenditures measured in US dollars adjusted for inflation and represented as a decimal.

4 Research Design

After consideration of the transit cost data and the research question, we decided to use a descriptive model to test the statistical power of time (years), tunnels, and stations relative to our dependent variable, realcost, on each transit project.

We started with EDA to remove records that did not have the necessary information (i.e. cost, # of tunnels, # of stations, location, etc.) to consider in the model testing. The records removed had null values in a critical dimension or metric field. We then created two columns to establish the length in time (years) for each project by calculating the difference between the end year and start year, and continent to investigate the dimensions with statistical significance relative to project cost in a region.

After cleansing the data, we used ggplot/geom_histogram to understand the distribution of the variables. We established that a log transform should be applied to give realcost a normal distribution. We then used ggplot/geom point to represent the linear relationship between realcost and potential independent variables. Variables considered were Length, TunnelDec, Stations, Elevated, Atgrade, and Tunnel.

After considering the plots we created linear models and performed z tests(coeftest in R) to understand the fit with regards to the p-value $< .05$ and overall statistical significance for each variable in each of the models. This approach was applied on the global dataset along with the Asia and Europe regions, which had enough sample size to pass I.I.D. assumptions.

Once the coeftest analysis was complete, we sought to understand fit. We built plots to represent the Residuals vs. Fitted, Normal Q-Q, Scale-Location, and Residuals vs. Leveraged. To test for bias, ggplots were built to compare the predicted vs residuals for each model. Once the bias within each model was established, we ran the stargazer function to calculate the R squared value to compare between each model.

Finally, we tested the CLM assumptions which were IID Data, No Perfect Collinearity, Linear Conditional Expectation, Homoskedastic Errors, and Normally Distributed Errors. Once the fit, bias, and assumptions were considered for each model, we decided on the best fitting model that had statistically significant variables, and high R squared values to explain the most variation in realcost. The coefficients for the independent variables were taken to establish which design elements best describe the total construction cost.

5 Model Building Process

In consideration of the model building process, we will want to measure the independent variables with the strongest statistical significance and highest explanation for the variation in our dependent variable (realcost) for the transit projects. Our goal is to understand the contributing features that impact realcost and understand if these features change when considering realcost for specific regions of the world. We decided to answer the research questions through 3 regional experiments: Global, Asia, and Europe.

5.1 Models At a Glance

We experimented different models on three distinct categories in the cleaned dataset: Global, Asia, and Europe. Ultimately, we made tradeoffs between statistical significance, collinearity between variables, and R-Squared values. At a glance, these are the models we concluded were the strongest for the particular data group.

5.1.1 Global

```
##
## Call:
## lm(formula = log(Realcost) ~ log(project_years) + log(Tunnel) +
##     log(Stations), data = projects)
##
## Coefficients:
##        (Intercept)  log(project_years)         log(Tunnel)       log(Stations)
##             5.1916              0.3343              0.4313              0.5114
```

Independent variables - project years, tunnel length, stations

Dependent variable - realcost

5.1.2 Asia

```
##
## Call:
## lm(formula = log(Realcost) ~ log(projectAS_years) + log(Tunnel) +
##     log(Stations) + log(Length), data = projects_AS)
##
## Coefficients:
##          (Intercept)  log(projectAS_years)           log(Tunnel)
##               5.1239                0.4752                0.1294
##        log(Stations)           log(Length)
##               0.2980                0.4565
```

Independent variables - project years, tunnel length, stations, overall length

Dependent variable - realcost

5.1.3 Europe

```
##
## Call:
## lm(formula = log(Realcost) ~ log(projectEU_years) + log(Tunnel),
##     data = projects_EU)
##
## Coefficients:
##          (Intercept)  log(projectEU_years)           log(Tunnel)
##               4.5827                0.4361                0.8753
```

Independent variables - project years, tunnel length

Dependent variable - realcost

5.2 Exploratory Data Analysis

By removing records with null values in a critical dimension, we are able to extract relationships between realcost and other variables. As we can see in the following plots, certain patterns and relations emerge.

Here are some notable observations:

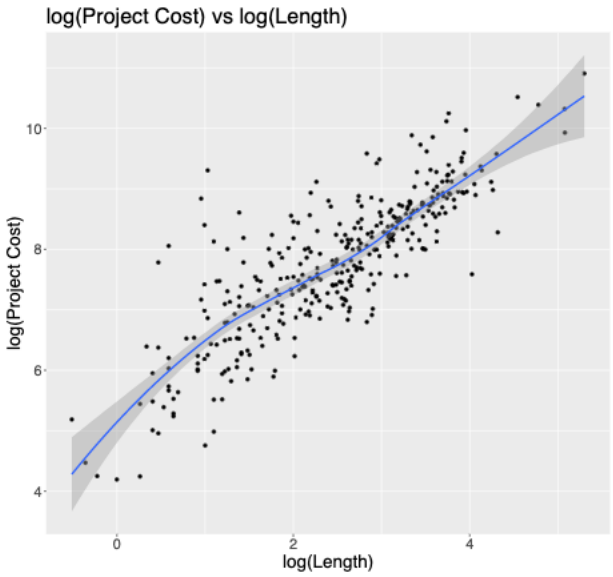
- log(Real cost) is normal.
- log(project years) is normal.
- log(length) and project cost is nearly linear, which is intuitive from an economics perspective because the longer an infrastructure takes, the more costly it is.

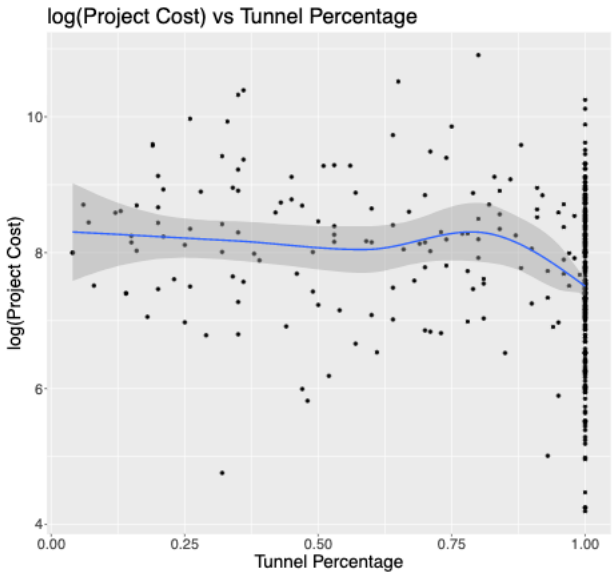
- log(stations) and project cost is linear because the more stations an infrastructure project has, the more expensive the project is to build the stations.
- Only Asia and Europ had over 50 projects in the dataset.

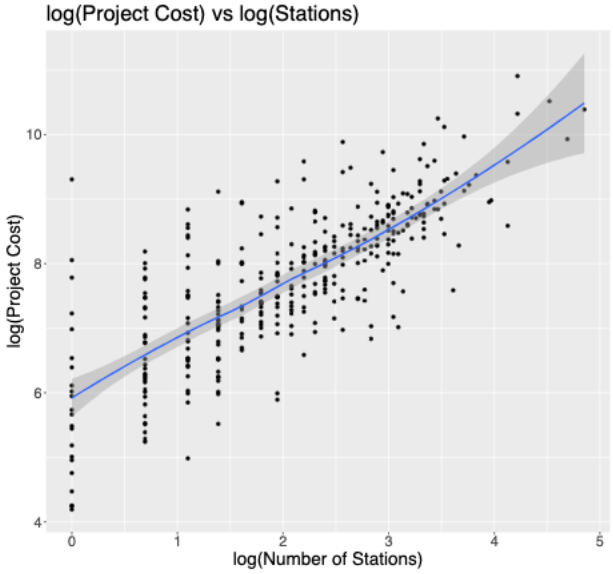
When looking at the dataset in aggregate, we're able to see normality in distribution of real cost and project years. However, the count of projects in each continent shows a different story and we see most of the infrastructure projects in the dataset are in Asia or Europe.

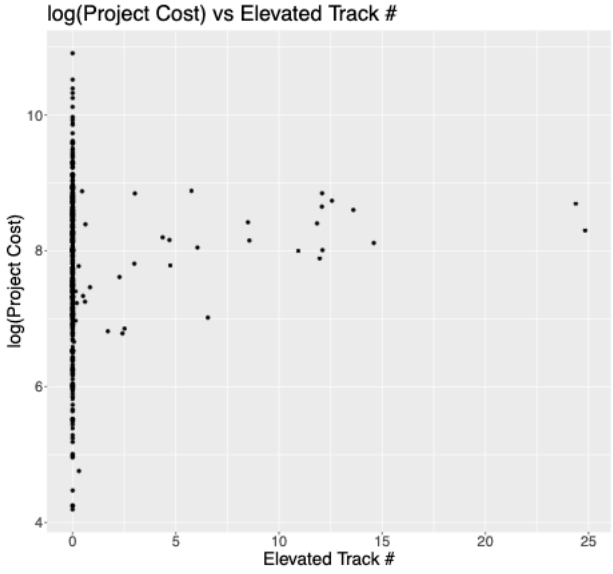
When plotting the linear relationship of each of these variables against realcost, the tunnel percentage, elevated track #, and atgrade did not show a clear relationship to cost or a trend that could be solved using a transform. Another problem was specifically in the collinearity that overall length has with stations and tunnel length variables. We chose not to use the overall length variable for the global model as the overall length diminishes the effect of the number of stations and tunnel length (model 6 compared to model 7 in the analysis). In the Asia experiment, collinearity was also the case for the overall length and tunnel length variables. However, we decided to keep overall length in the end as tunnel length was still statistically significant still and the model with overall length explained the highest amount of variation in realcost at .805. The Europe experiment had the strongest effect of collinearity between tunnel length and overall length/stations. This is why the model for the Europe experiment is best fit with project years and tunnel length specifically.

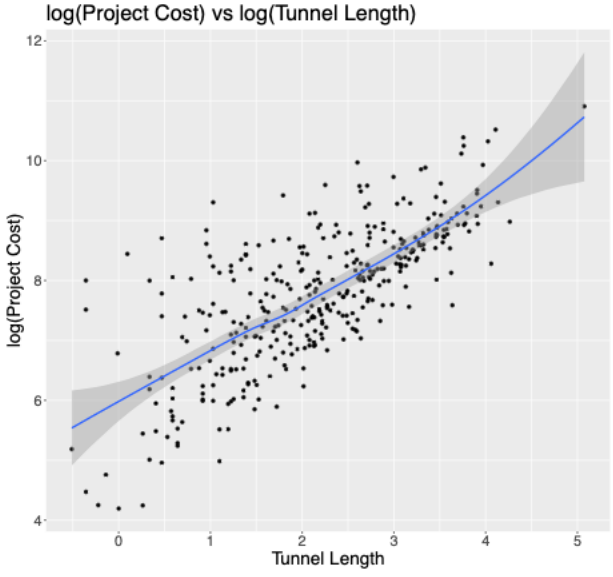
In each case in order to achieve a normal distribution and a linear relationship to realcost a log transform had to be performed on each of the variables. Specifically, log(Realcost), log(project_years), log(Stations), log(Tunnel), and log(Length). This was revealed when plotting the density/linear relationships of each variable and correlation to realcost. The transform decision was also clear when running coeftest/stargazer to determine the statistical significance and coefficients for each model fit when considering the variables.


log(Project Cost) vs log(Length)
log(Project Cost)
log(Length)


log(Project Cost) vs Tunnel Percentage
log(Project Cost)
Tunnel Percentage


log(Project Cost) vs log(Stations)
log(Project Cost)
log(Number of Stations)


log(Project Cost) vs Elevated Track #
log(Project Cost)
Elevated Track #


log(Project Cost) vs log(Tunnel Length)
log(Project Cost)
Tunnel Length


log(Project Cost) vs project duration
log(Project Cost)
Project duration in years

5.3 Building the Models

To understand the covariates that would help us decide which of the contributing features would impact the cost of a rapid transit project, we first performed EDA (exploratory data analysis). The variables considered included the overall length, tunnel percentage, tunnel length, stations, elevated track #, years to complete, and atgrade for each transit project.

Each of our choices for EDA was supported by visual tools or statistical tests. We decided to remove nearly 300 records from the dataset as a data cleansing step after seeing that there were many variables with null values that would cause bias in the distribution and statistical power in the model fit.

We created a continent field to compare regions because the sample size was too small for many of the countries to perform any country specific regression analysis. We also built the project years field by subtracting the end year and start year of each transit project.

Visually we used ggplot to represent the distribution and relationships of each proposed independent variable to cost. This allowed us to understand which variables did not show a relationship to cost or would require transformations to achieve a normal distribution, as shown in the Exploratory Data Analysis section.

We then built separate models testing the fit for each combination of variables using stargazer. We supported this with separate coeftest to see the specific p value for each independent variable. We followed this step with plots showing the Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage relationships to test for fit. When testing for bias we compared the predicted vs residuals for the models, looking for a relatively flat trend line in the relationship. Last, the 5 CLM assumptions were tested to ensure we did not miss violations for I.I.D., Perfect Collinearity, Linear Conditional Expectation, Homoskedastic Errors, or Normally Distributed Errors.

5.4 Creating a Global Model

The global model uses all projects as datapoints from the dataset. We implemented seven different models and compared them against each other, using R-Squared and feature statistical significance as the gauge to determine the best estimator.

> Note: Below in the model equations, the bolded model is the best model selected out of the Global models.

5.4.1 Global Models

$GL1 = log(Realcost) = \beta_0 + \beta_1 log(project\_years)$

$GL2 = log(Realcost) = \beta_0 + \beta_1 log(Tunnel)$

$GL3 = log(Realcost) = \beta_0 + \beta_1 log(Stations)$

$GL4 = log(Realcost) = \beta_0 + \beta_1 log(Length)$

$GL5 = log(Realcost) = \beta_0 + \beta_1 log(project\_years) + \beta_2 log(Tunnel)$

$\mathbf{GL6} = log(Realcost) = \beta_0 + \beta_1 log(project\_years) + \beta_2 log(Tunnel) + \beta_3 log(Stations)$

$GL7 = log(Realcost) = \beta_0 + \beta_1 log(project\_years) + \beta_2 log(Tunnel) + \beta_3 log(Stations + 1) + \beta_4 log(Length)$

5.5 Creating Continental Models - Asia and Europe

Since only Asia and Europe had over 100 data points, we built models specific to these continents. The continental models focused only on projects in Europe and Asia separately. Seven Asia models and nine Europe models were implemented and compared against each other, similar to the Global Models.

> Note: Below in the model equations, the bolded model is the best model selected out of the Asia and Europe models.

5.5.1 Asia Models

$AS1 = log(Realcost) = \beta_0 + \beta_1 log(projectAS\_years)$

$AS2 = log(Realcost) = \beta_0 + \beta_1 log(Tunnel)$

$AS3 = log(Realcost) = \beta_0 + \beta_1 log(Stations)$

$AS4 = log(Realcost) = \beta_0 + \beta_1 log(Length)$

$AS5 = log(Realcost) = \beta_0 + \beta_1 log(projectAS\_years) + \beta_2 log(Tunnel)$

$AS6 = log(Realcost) = \beta_0 + \beta_1 log(projectAS\_years) + \beta_2 log(Tunnel) + \beta_3 log(Stations)$

$\mathbf{AS7} = log(Realcost) = \beta_0 + \beta_1 log(projectAS\_years) + \beta_2 log(Tunnel) + \beta_3 log(Stations) + \beta_4 log(Length)$

5.5.2 Europe Models

$EU1 = log(Realcost) = \beta_0 + \beta_1 log(projectEU\_years)$

$EU2 = log(Realcost) = \beta_0 + \beta_1 log(Tunnel)$

$EU3 = log(Realcost) = \beta_0 + \beta_1 log(Stations)$

$EU4 = log(Realcost) = \beta_0 + \beta_1 log(Length)$

$\mathbf{EU5} = log(Realcost) = \beta_0 + \beta_1 log(projectEU\_years) + \beta_2 log(Tunnel)$

$EU6 = log(Realcost) = \beta_0 + \beta_1 log(projectEU\_years) + \beta_3 log(Stations)$

$EU7 = log(Realcost) = \beta_0 + \beta_1 log(projectEU\_years) + \beta_2 log(Tunnel) + \beta_3 log(Stations)$

$EU8 = log(Realcost) = \beta_0 + \beta_1 log(projectEU\_years) + \beta_2 log(Tunnel) + \beta_3 log(Length)$

$EU9 = log(Realcost) = \beta_0 + \beta_1 log(projectEU\_years) + \beta_2 log(Tunnel) + \beta_3 log(Stations) + \beta_4 log(Length)$

6 Testing the 5 Classical Linear Model Assumptions

6.1 Testing 5 CLM Assumptions - Global Model

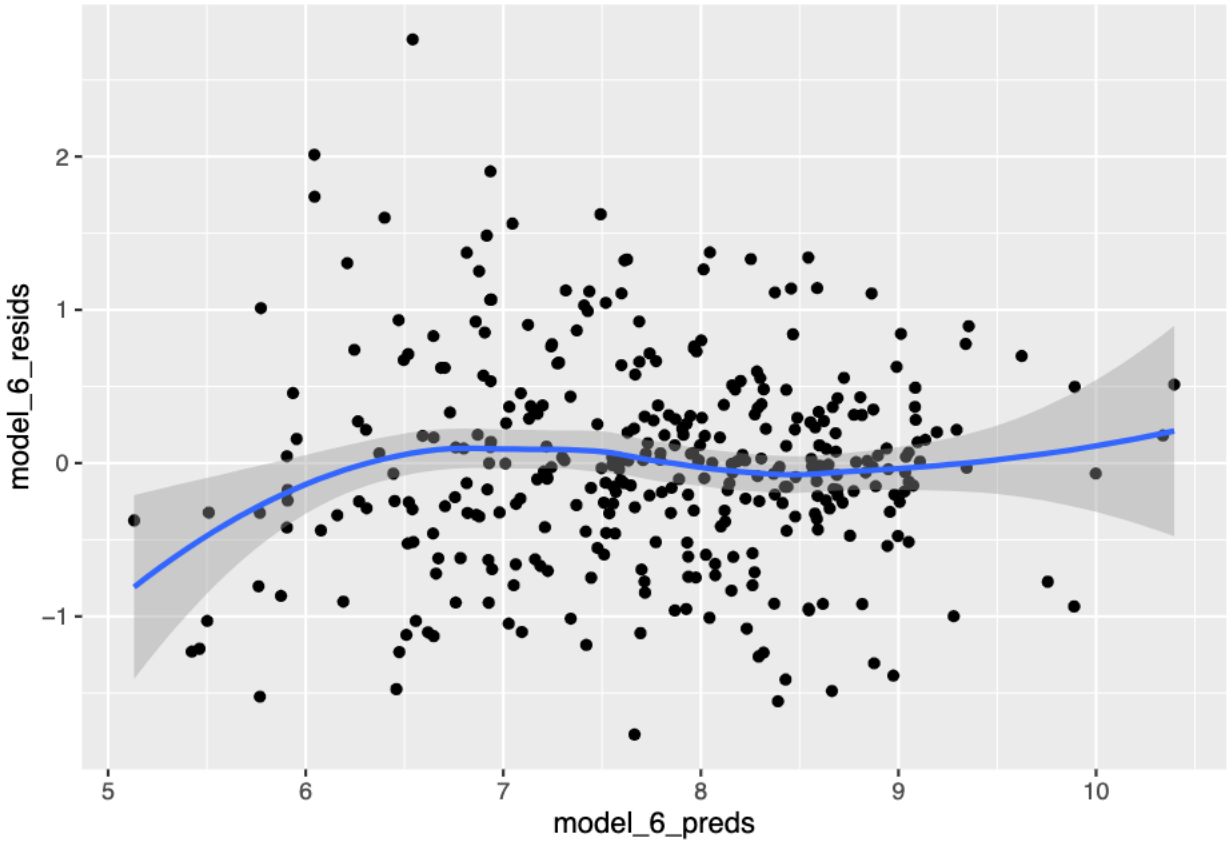
1. IID Sampling

   This dataset doesn't meet IID requirements, becuase of the following:

   The data was collected from many different sources; however, the researchers stated that they prefered to use the data from the most recent source. This could undermine random sampling because data from some certain countries might have more recent and accurate information than the other regions.

   Also, the projects in same countries and continents tend to show similar project costs which shows clustering.

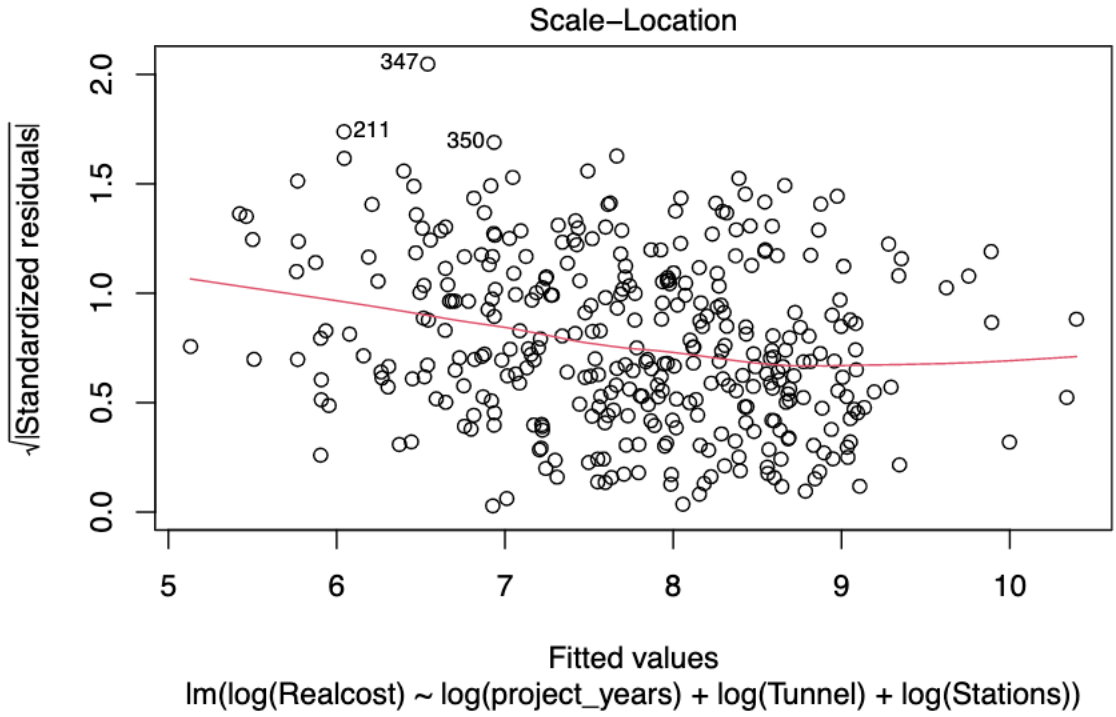
2. Linear Conditional Expectation



The model shows no obvious residuals' deviation from 0 on both side of the x-axis. This test confirms that the linear condition expectation is met in this model. 3. No Perfect Collinearity

```
##       (Intercept) log(project_years)       log(Tunnel)      log(Stations)
##         5.1915758          0.3343008          0.4312727          0.5113656
```

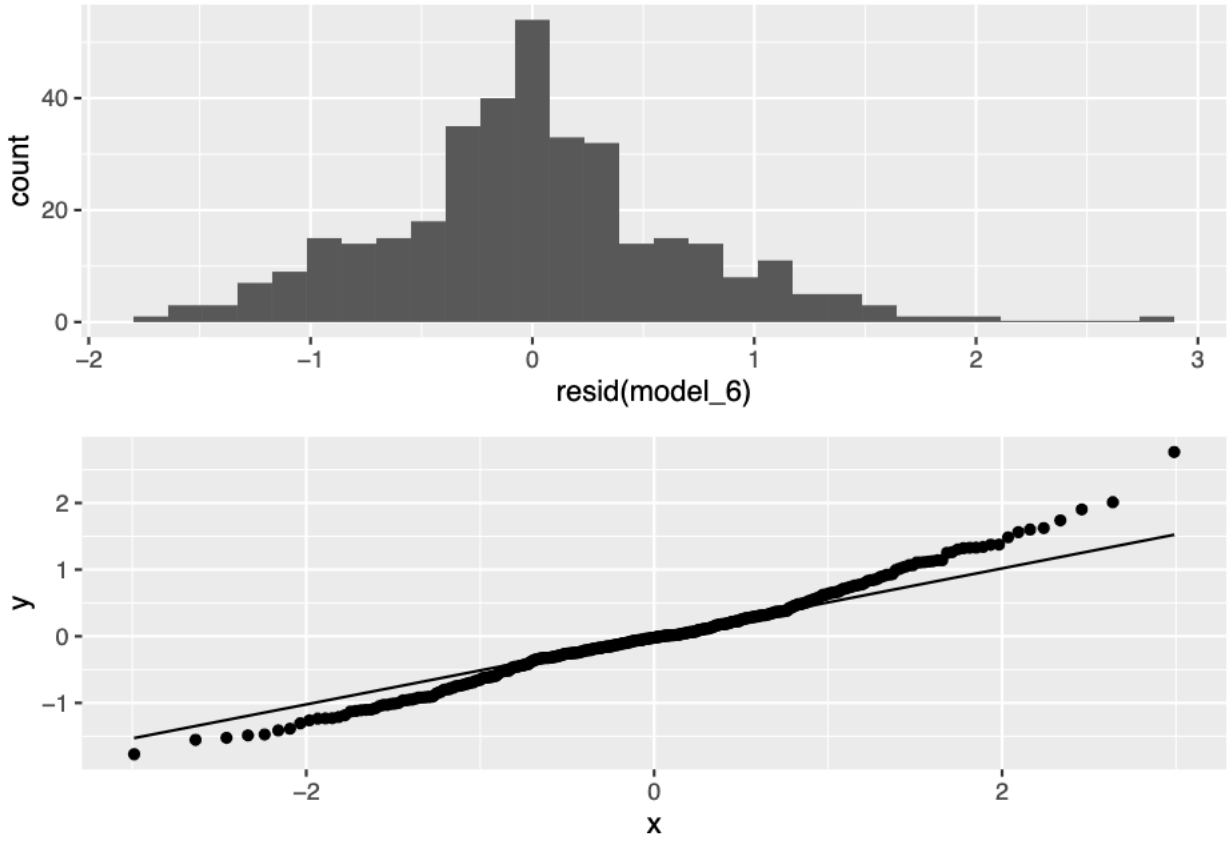
By looking at the model coefficient we see that R hasn't dropped any of the variables. This tells us that there is no perfect collinearity. This assumption also includes the requirement that a BLP exists, which may not happen if there are heavy tails. In this case, though, we don't see any distributions that look like they have unusually low or high values.



4. Homoskedastic Errors

   Using the scale-location plot, homoskedasticity should show up on this plot as a flat smoothing curve; however, the plot shows that on the left side of the plot the residuals are spreading wider, but as the model progress to the right side the residuals starts to get closer to a normal distribution and the line becomes horizontal. Despite what is stated above, because the angle is not steep, it suggests there is no major problem with homoskedasticity.

5. Normally Distributed Errors



Although there is a slight deviation on both side of the Q-Q plot. The residuals distribution matches a normal distribution without any major deviations from normality. The only concern would be the data point on the top right side of the Q-Q plot that shows to be a little off.

6.2 Testing 5 CLM Assumptions - Asia Model

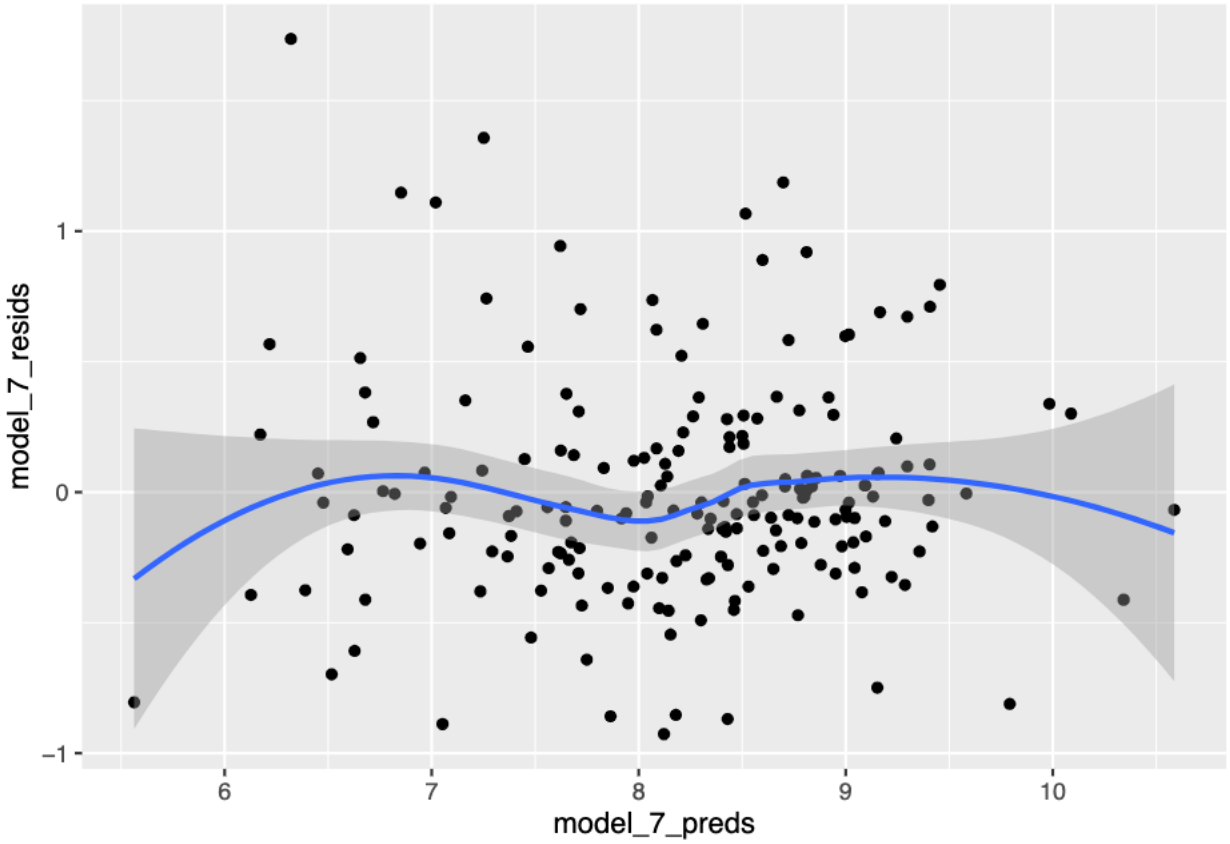
1. IID Sampling

   This dataset doesn't meet IID requirements, becuase of the following:

   The data was collected from many different sources; however, the researchers stated that they prefered to use the data from the most recent source. This could undermine random sampling because data from some certain countries might have more recent and accurate information than the other regions.

   Also, the projects in same countries tend to show similar project costs which shows clustering.

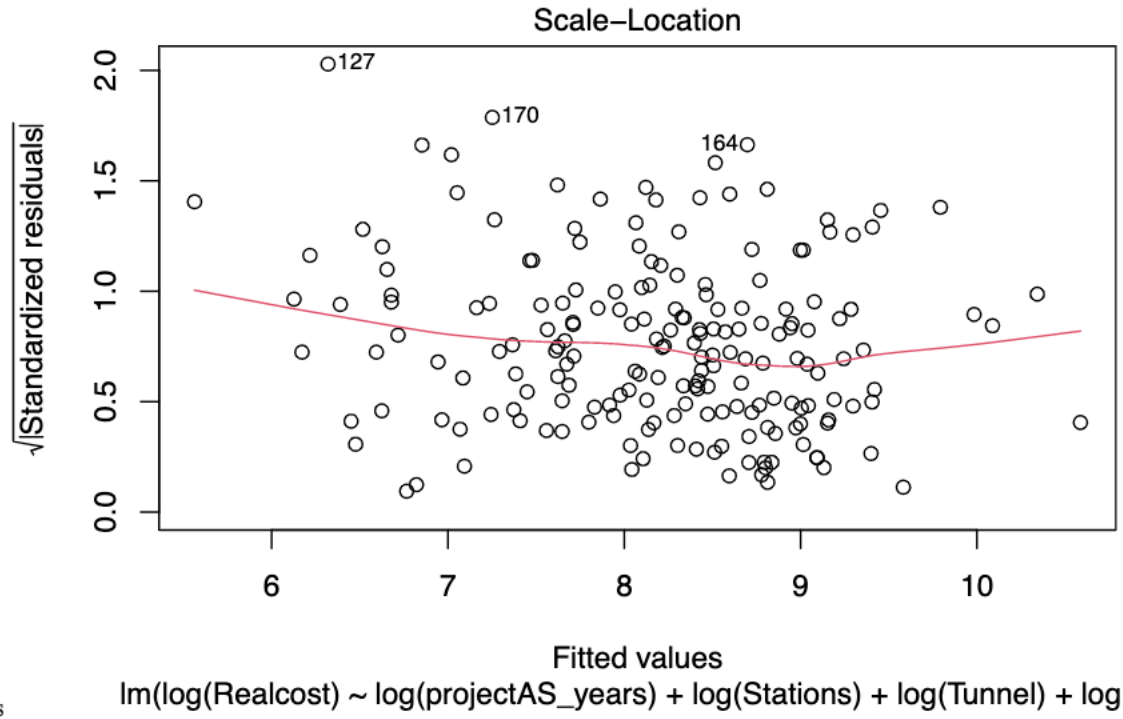
2. Linear Conditional Expectation



> The model shows no obvious residuals' deviation from 0 on both side of the x-axis. This test confirms that the linear condition expectation is met in this model. 3. No Perfect Collinearity

```
##          (Intercept) log(projectAS_years)        log(Stations)
##            5.1238944            0.4751883            0.2980294
##         log(Tunnel)          log(Length)
##            0.1293860            0.4564575
```

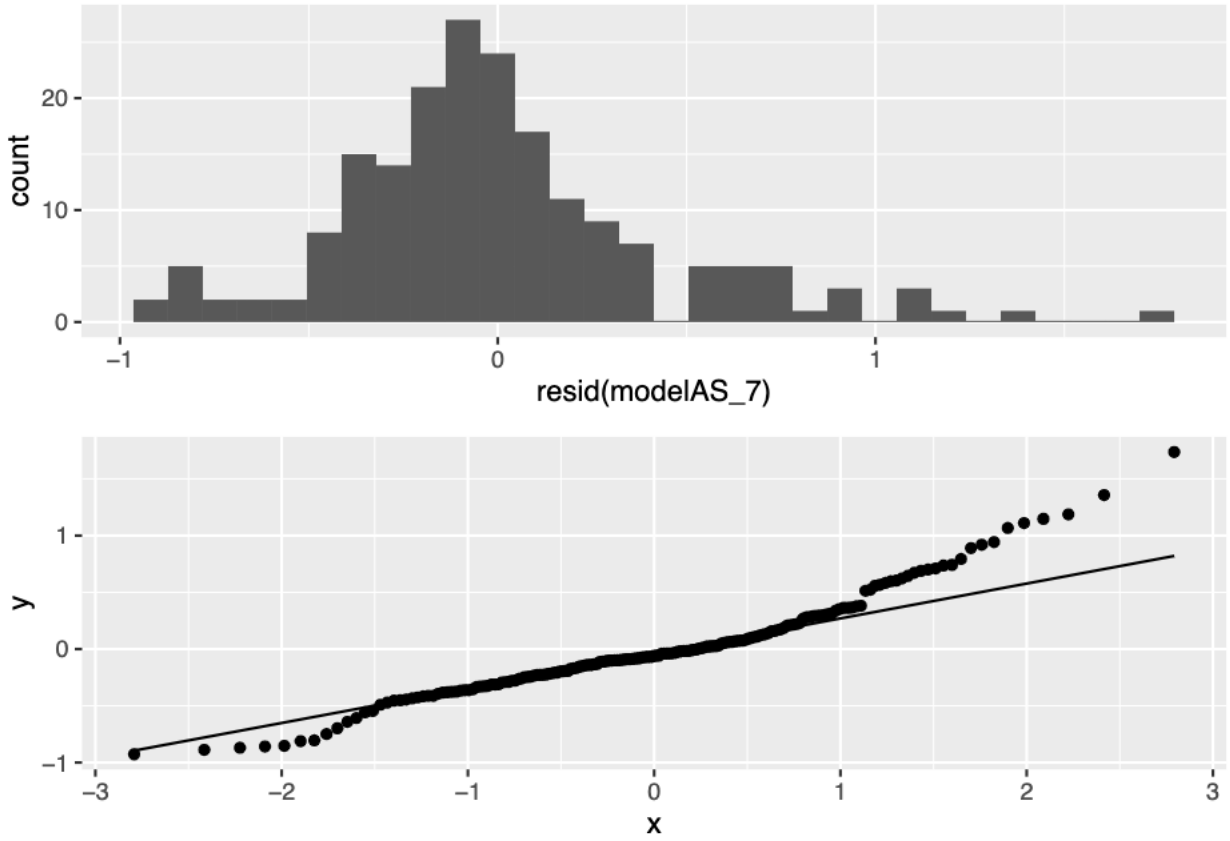
   By looking at the model coefficient we see that R hasn't dropped any of the variables. This tells us that there is no perfect collinearity. This assumption also includes the requirement that a BLP exists, which may not happen if there are heavy tails. In this case, though, we don't see any distributions that look like they have unusually low or high values.



4. Homoskedastic Errors

   Using the scale-location plot, homoskedasticity should show up on this plot as a flat smoothing curve. The above plot shows an almost horizontal line which suggests that there is no major problem with heteroskedasticity.

5. Normally Distributed Errors



In both plots, the residual distribution shows a strong deviation from normality on the right side of the plot. This would reject the normally distributed errors requirement of CLM; however, because of the large sample size and central limit theorem we can say the the residuals are asymptotically normally distributed.

6.3 Verifying 5 CLM Assumptions - Europe Model

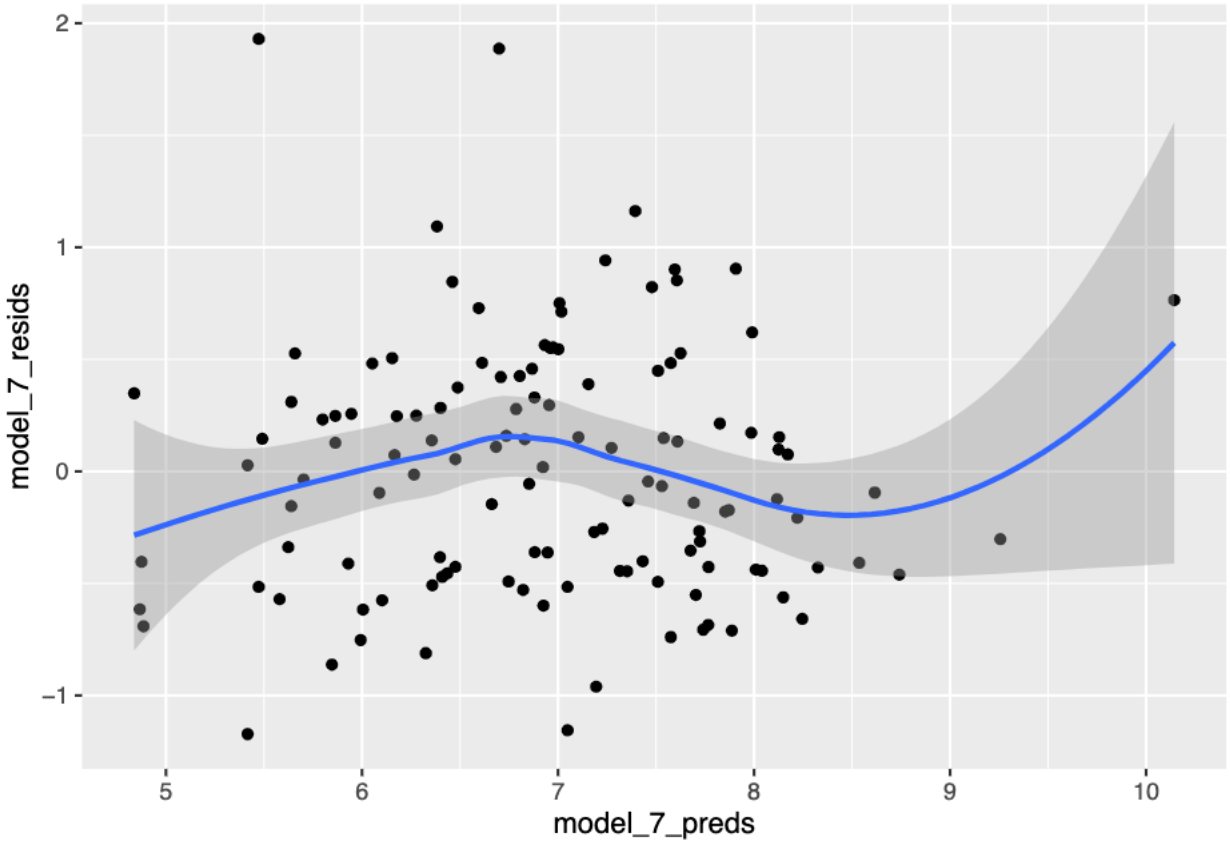
1. IID Sampling

   This dataset doesn't meet IID requirements, becuase of the following:

   The data was collected from many different sources; however, the researchers stated that they prefered to use the data from the most recent source. This could undermine random sampling because data from some certain countries might have more recent and accurate information than the other regions.

   Also, the projects in same countries tend to show similar project costs which shows clustering.

2. Linear Conditional Expectation

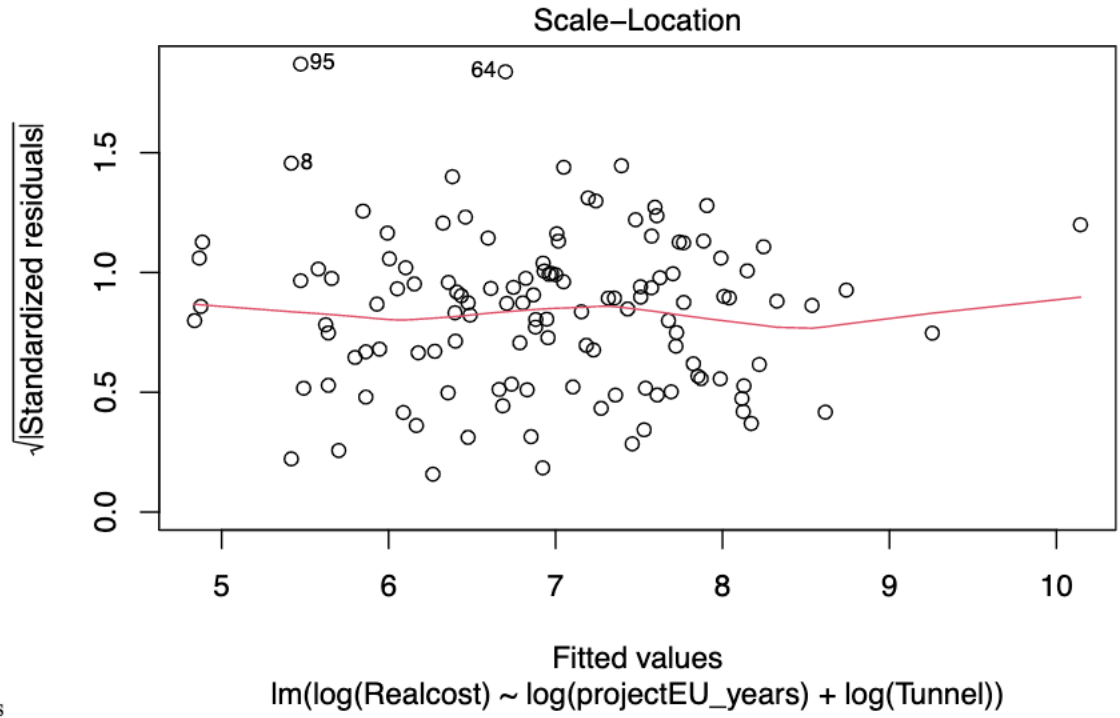


   The model shows no obvious residuals' deviation from 0. This test confirms that the linear condition expectation is met in this model.

3. No Perfect Collinearity

```
##          (Intercept) log(projectEU_years)          log(Tunnel)
##            4.5827386            0.4360791            0.8752879
```

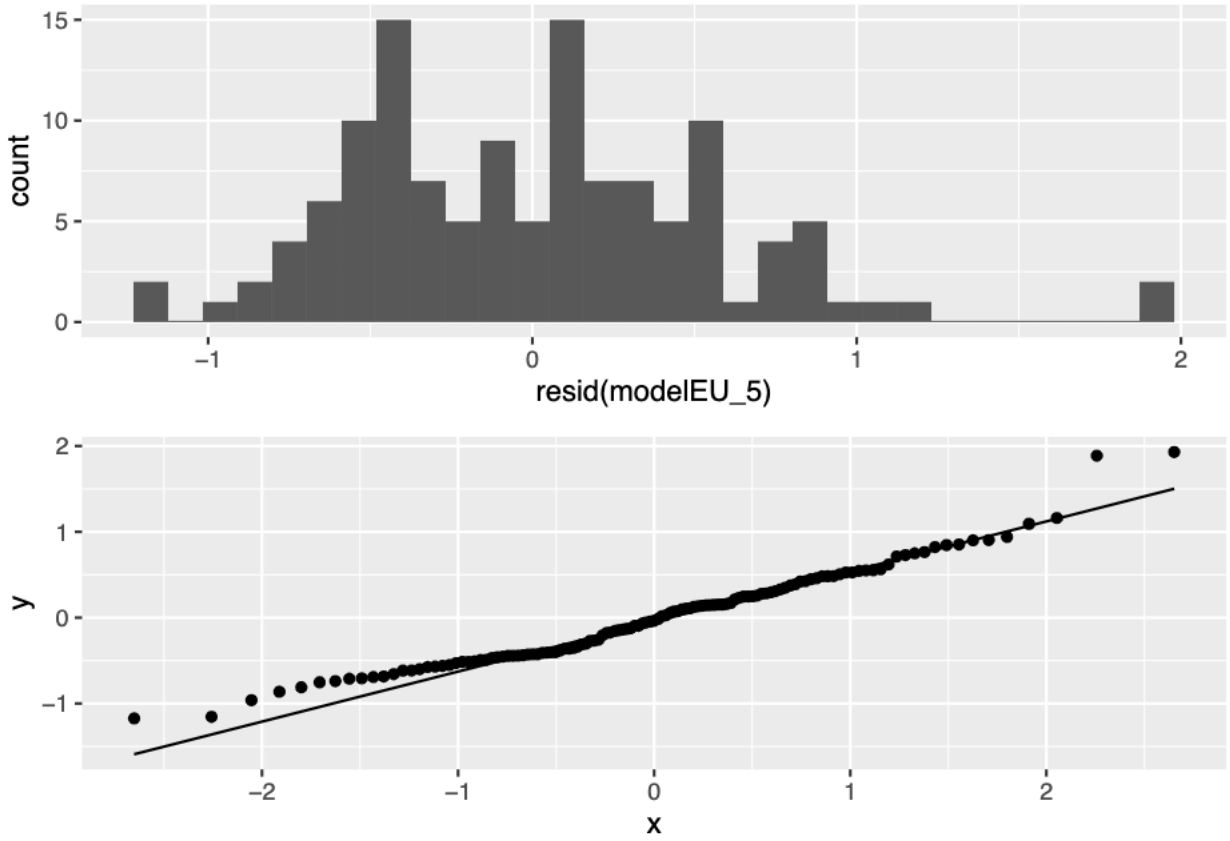
   By looking at the model coefficient we see that R hasn't dropped any of the variables. This tells us that there is no perfect collinearity. This assumption also includes the requirement that a BLP exists, which may not happen if there are heavy tails. In this case, though, we don't see any distributions that look like they have unusually low or high values.



4. Homoskedastic Errors

   Using the scale-location plot, homoskedasticity should show up on this plot as a flat smoothing curve. The above plot shows an almost horizontal line which suggests that there is no major problem with heteroskedasticity.

5. Normally Distributed Errors



The Q-Q plot and resisual histogram, show a normal residual distribution that meets the normally distributed errors requirement of CLM.

7 Results and Comparing the Chosen Models

Overall, we see that project years and tunnels play a significant role in the Global, Asia, and Europe models, contributing to the cost of a rapid transit project.

By comparing the coefficients and these two variables appearing in all three developed models, we see the log(project years) in different magnitudes in all the models, showing variance is emerging from something else not accounted for in the data. The t-test coefficients show us the statistical significance is .001. Intuitively, this also makes sense because projects that span multiple years may result in complexities that increase the cost of a rapid transit project. However, because of how different this coefficient was across the different models, we also see regional differences that make cost structures so different.

Further, tunnels, stations, and length all seem to be proxies for each other, as they would all be positively correlated from an economic point of view; the more tunnels and stations an infrastructure project would have, presumably the length would be longer to fit all the tunnels and stations. As a result, we cannot fully determine to what extent tunnels, stations and length cause differences in the cost of rapid transit projects across the different continents in our experiments. Moreover, because the t-test of these coefficients all show .001 statistical significance, we can infer that geographic locations do in fact cause a significant difference in cost.

7.1 Global Model

We chose the log transformation for independent and dependent variables, because first, we don't think there is a linear relationship between them. Second, all of the variables have a positive, non-zero distribution, and the log transformation made their distribution more symmetric and normal.

In model 6, all three coefficients are statistically significant so we can conclude that there is a statistically meaningful relationship between log transformation of real cost and duration of project, length of tunnel, and number of stations. The coefficients of log(project_years), log(Tunnel), and log(Stations) are all statistically significant and positive, which means an increase in any of the dependent variables is associated with an increase in real cost at an increasing rate.

According to the regression table, project_years is statistically significant, and as the number of predictors included in the model 6 increases, the estimated relationship between real cost and education tends to become less positive. This is evidence of the robustness of our model 6. The table also shows that the adjusted R^2 is increasing which means that the model is becoming better in explaining the variance of the data.

Our results show that the addition of the length variable diminishes the effect of the number of stations and that shows there is a collinearity between the length and number of stations.

7.2 Asia Model

In modelAS_7, all four coefficients are statistically significant so we can conclude that there is a statistically meaningful relationship between log transformation of real cost and duration of project, length of tunnel, and number of stations. The coefficients of log(project_years), log(Tunnel), log(Stations), log(length) are all statistically significant and positive, which means an increase in any of the dependent variables is associated with an increase in real cost at an increasing rate.

According to the regression table, project_years is statistically significant, and as the number of predictors included in the model 6 increases, the estimated relationship between real cost and project years tends to become less positive. This is evidence of the robustness of our model 6. The table also shows that the adjusted R^2 is increasing which means that the model is becoming better in explaining the variance of the data.

7.3 Europe Model

In modelEU_5, both coefficients are statistically significant so we can conclude that there is a statistically meaningful relationship between log transformation of real cost and duration of project, length of tunnel. Both coefficients of log(project_years) and log(Tunnel) are statistically significant and positive, which means an increase in any of the two variables is associated with an increase in real cost at an increasing rate.

According to the regression table, project_years is statistically significant, and as the number of predictors included in the model 5 increases, the estimated relationship between real cost and project_years tends to become less positive. This is evidence of the robustness of our model 5. The table also shows that the adjusted R^2 is increasing which means that the model is becoming better in explaining the variance of the data.

Our results show that the addition of both number of stations and length variables diminishes the effect of number of tunnel lengths and that shows there is a collinearity between the length, number of stations and tunnel_length.

Here, the model coefficients are shown:

Table 1:

	Dependent variable:		
	log(Realcost)		
	GL6	AS7	EU5
	(1)	(2)	(3)
log(project_years)	0.334***		
	(0.093)		
log(projectAS_years)		0.475***	
		(0.103)	
log(projectEU_years)			0.436***
			(0.105)
log(Tunnel)	0.431***	0.129**	0.875***
	(0.064)	(0.049)	(0.057)
log(Length)		0.456***	
		(0.105)	
log(Stations)	0.511***	0.298***	
	(0.067)	(0.090)	
Constant	5.192***	5.124***	4.583***
	(0.168)	(0.176)	(0.210)
Observations	358	191	125
R^2	0.668	0.809	0.745
Adjusted R^2	0.665	0.805	0.741
Residual Std. Error	0.671 (df = 354)	0.432 (df = 186)	0.561 (df = 122)
F Statistic	237.179*** (df = 3; 354)	196.936*** (df = 4; 186)	178.395*** (df = 2; 122)
Note:			*p<0.05; **p<0.01; ***p<0.001

8 Limitations of the Model

8.1 Statistical Limitations

- The sample data only included all transit projects for which public information was available through a variety of sources.
- Large samples were available for Asia and Europe and N.America, S.America and ANZ samples were much smaller.
- Collinearity between length and real cost, length with stations and tunnel length.
- Needed transformation (log) to ensure normal distribution.

8.2 Strutural Limitations

- In-tangible costs like quality of project management, availability of skilled resources and local politics have major bearings on project costs.
- As an example in reviewing the Green Line Extension in Boston, much of the cost increases were attributed to change from a Democratic Governor to a Republican Governor, cancellation of the project and laying off critical project management personnel.
- Large number of records did not have reliable data.

9 Conclusion

As shown, we see how different the optimal model and model coefficients have become based on the data alone. This shows the regional differences that motivated this study. We found statistical significance in all the transformed features we've chosen through our experiments in the dataset.

In conclusion, our key finding shows that project duration and tunnel length have the strongest support for impacting the cost of a rapid transit project. Additionally, while we found statistically significant support for the other features, such as stations and project length, the variability in the coefficients points to variance in the datasets used, which may indicate significant differences in costs due to the geographic location of a project.

The implications for this study shows that projects should minimize tunnel length and project duration to have cost-effective transit systems.

10 Appendix (for reference only)

10.1 References

1. Transit Cost Project - We are a group of researchers under the umbrella of the NYU Marron Institut
2. Transit Cost Dataset - (https://transitcosts.com/data/)
3. The Boston Case: The Story of the Green Line Extension - (https://transitcosts.com/city/boston-cas
4. Istanbul M1B - LA Purple Line Comparison - (https://transitcosts.com/city/istanbul-m1b-la-purple-l

10.2 Best Model Coefficients

```
## [1] "Global Model"
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         5.191576   0.167522 30.9904 < 2.2e-16 ***
## log(project_years) 0.334301   0.093309  3.5827 0.0003876 ***
## log(Tunnel)         0.431273   0.063666  6.7740 5.235e-11 ***
## log(Stations)       0.511366   0.067164  7.6137 2.446e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## [1] "Asia Model"
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          5.123894   0.176212 29.0779 < 2.2e-16 ***
## log(projectAS_years) 0.475188   0.102826  4.6213 7.112e-06 ***
## log(Stations)        0.298029   0.089988  3.3119  0.001113 **
## log(Tunnel)          0.129386   0.049004  2.6403  0.008987 **
## log(Length)          0.456457   0.105077  4.3440 2.297e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## [1] "Europe Model"
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          4.582739   0.209712 21.8525 < 2.2e-16 ***
## log(projectEU_years) 0.436079   0.105418  4.1367 6.508e-05 ***
## log(Tunnel)          0.875288   0.056792 15.4121 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10.3 Model Tables

Note: We choose model 6 over model 7, because addition of length caused collinearity which diminishes the effect of number of stations.

Note we choose model 7 over model 6, because the length does not show collinearity with number of stations when we look at Asia data.

Table 2:

	Dependent variable:			
	log(Realcost)			
	GL1	GL2	GL3	GL4
	(1)	(2)	(3)	(4)
log(project_years)	0.575***			
	(0.133)			
log(Tunnel)		0.838***		
		(0.050)		
log(Stations)			0.868***	
			(0.046)	
log(Length)				0.927***
				(0.038)
Constant	6.763***	5.956***	5.955***	5.487***
	(0.242)	(0.132)	(0.118)	(0.111)
Observations	358	358	358	358
R^2	0.058	0.548	0.597	0.700
Adjusted R^2	0.056	0.547	0.596	0.699
Residual Std. Error (df = 356)	1.126	0.780	0.737	0.636
F Statistic (df = 1; 356)	22.036***	432.064***	526.916***	828.993***
Note:				*p<0.05; **p<0.01; ***p<0.001

Table 3:

	Dependent variable:		
	log(Realcost)		
	GL5	GL6	GL7
	(1)	(2)	(3)
log(project_years)	0.467***	0.334***	0.411***
	(0.087)	(0.093)	(0.088)
log(Tunnel)	0.824***	0.431***	0.131**
	(0.047)	(0.064)	(0.050)
log(Stations)		0.511***	
		(0.067)	
log(Stations + 1)			0.005
			(0.086)
log(Length)			0.798***
			(0.081)
Constant	5.180***	5.192***	4.802***
	(0.187)	(0.168)	(0.149)
Observations	358	358	358
R^2	0.587	0.668	0.733
Adjusted R^2	0.584	0.665	0.730
Residual Std. Error	0.747 (df = 355)	0.671 (df = 354)	0.603 (df = 353)
F Statistic	251.809*** (df = 2; 355)	237.179*** (df = 3; 354)	241.725*** (df = 4; 353)
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 4:

	Dependent variable:			
	log(Realcost)			
	AS1	AS2	AS3	AS4
	(1)	(2)	(3)	(4)
log(projectAS_years)	0.923***			
	(0.178)			
log(Tunnel)		0.713***		
		(0.066)		
log(Stations)			0.854***	
			(0.049)	
log(Length)				0.905***
				(0.052)
Constant	6.745***	6.431***	6.166***	5.624***
	(0.285)	(0.190)	(0.137)	(0.168)
Observations	191	191	191	191
R^2	0.172	0.522	0.724	0.740
Adjusted R^2	0.168	0.520	0.722	0.739
Residual Std. Error (df = 189)	0.892	0.678	0.515	0.500
F Statistic (df = 1; 189)	39.268***	206.746***	495.308***	538.144***
Note:				*p<0.05; **p<0.01; ***p<0.001

Table 5:

	Dependent variable:		
	log(Realcost)		
	AS5	AS6	AS7
	(1)	(2)	(3)
log(projectAS_years)	0.735***	0.466***	0.475***
	(0.111)	(0.103)	(0.103)
log(Stations)		0.605***	0.298***
		(0.063)	(0.090)
log(Tunnel)	0.672***	0.256***	0.129**
	(0.058)	(0.056)	(0.049)
log(Length)			0.456***
			(0.105)
Constant	5.375***	5.393***	5.124***
	(0.213)	(0.167)	(0.176)
Observations	191	191	191
R^2	0.630	0.787	0.809
Adjusted R^2	0.626	0.783	0.805
Residual Std. Error	0.598 (df = 188)	0.455 (df = 187)	0.432 (df = 186)
F Statistic	159.962*** (df = 2; 188)	229.643*** (df = 3; 187)	196.936*** (df = 4; 186)
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 6:

	Dependent variable:		
	log(Realcost)		
	EU1	EU2	EU3
	(1)	(2)	(3)
log(projectEU_years)	1.046***		
	(0.176)		
log(Tunnel)		0.962***	
		(0.062)	
log(Stations)			0.834***
			(0.075)
Constant	4.989***	5.245***	5.545***
	(0.349)	(0.127)	(0.152)
Observations	125	125	125
R^2	0.239	0.709	0.589
Adjusted R^2	0.232	0.707	0.586
Residual Std. Error (df = 123)	0.966	0.597	0.710
F Statistic (df = 1; 123)	38.529***	300.300***	176.397***
Note:			*p<0.05; **p<0.01; ***p<0.001

Note: We should choose model 5, because addition of number of stations and length has collinearity with total tunnel length. While model 8 has the higher R^2, we believe the strong corelation between length and cost is a concern for data leakage.

Table 7:

	Dependent variable:		
	log(Realcost)		
	EU4	EU5	EU6
	(1)	(2)	(3)
log(Length)	0.926***		
	(0.057)		
log(projectEU_years)		0.436***	0.431**
		(0.105)	(0.143)
log(Tunnel)		0.875***	
		(0.057)	
log(Stations)			0.742***
			(0.076)
Constant	5.181***	4.583***	4.894***
	(0.118)	(0.210)	(0.260)
Observations	125	125	125
R^2	0.736	0.745	0.623
Adjusted R^2	0.733	0.741	0.616
Residual Std. Error	0.569 (df = 123)	0.561 (df = 122)	0.683 (df = 122)
F Statistic	342.090*** (df = 1; 123)	178.395*** (df = 2; 122)	100.634*** (df = 2; 122)
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 8:

	Dependent variable:		
	log(Realcost)		
	EU7	EU8	EU9
	(1)	(2)	(3)
log(projectEU_years)	0.391*	0.417	0.420
	(0.176)		
log(Tunnel)	0.734	0.295***	0.295
		(0.062)	
log(Stations)	0.173		−0.013
			(0.075)
log(Length)		0.587	0.597
Constant	4.626***	4.527***	4.523***
	(0.349)	(0.127)	(0.152)
Observations	125	125	125
R^2	0.753	0.779	0.779
Adjusted R^2	0.747	0.773	0.772
Residual Std. Error	0.555 (df = 121)	0.525 (df = 121)	0.527 (df = 120)
F Statistic	122.869*** (df = 3; 121)	142.145*** (df = 3; 121)	105.747*** (df = 4; 120)
Note:			*p<0.05; **p<0.01; ***p<0.001