Annex 7: Statistical analysis on unit rail costs

A statistical approach was undertaken to assess the impact of the various factors identified on the level of unit cost in the railway industry.

The following hypothesis is tested:

- The length and composition of the infrastructure produce an impact on the cost of the infrastructure.

To test for this hypothesis, statistical techniques have been used with the aim to isolate the statistical relationships between one variable (the cost of the infrastructure) with multiple factors that identify the complexities relative to the infrastructure type, orography, etc.

1.1. High-level methodology

To test the statistical relevance of the factors supposed to impact on the cost of rail infrastructure, a regression analysis is performed.

Hereunder, the table reports the dependent variables considered in this testing.

Variable	Description	Units	Source
Total cost	Total cost of the project included in the database	M€	Database
Construction cost	Cost of the project relative to the construction works	M€	Database
PPP factor	Purchasing power parity	€ local / € EU28 (2017)	Eurostat
Inflation factor		curr. / curr.(2017)	Eurostat
Length	The length of the infrastructure is defined as the length in terms of double-track equivalent	Km	Database
Category	The category variable defines the type of investment type that is included in the database.	Type of investment: • New line; • Rehabilitation and upgrade; • Signalling, electrification; • Big infrastructure (tunnels, stations, hubs, etc)	Database
Infrastructure type		High-speed / conventional	Database
Tunnels total length	The length of the tunnels on the line	Km	Database
Bridges total length	The length of the bridges on the line	Km	Database
Viaduct total length	The length of the viaducts on the line	Km	Database
Number of interfaces		Number	Database
Number of stations		Number	Database
Base year	Year on which the		Database

	information on costs are relative to		
Predominant environment	Type of the environment that is crossed by the line	Rural/ Urban	Database
Terrain description	Type of terrain that is crossed by the line	Flat/ Hilly/ Mountaneous	Database
Energy	Type of propellant used to operate on the line	Electric/ Diesel	Database
Design Speed	Design speed of the line	Km/h	Database

1.2. Multiple regression analysis

To estimate the elements impacting on the rail unit cost and their relative impact, different models are used. The multiple regression analysis is used to predict the value of one variable from a set of predictors.

For the purpose of our analysis, the central hypothesis is that

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \textit{(Equation 1)}$$

Where **y** is the vector of infrastructure cost, $\boldsymbol{X}$ is the design matrix, which represents the elements that impact on the cost of the infrastructure, $\boldsymbol{\beta}$ is the vector of the coefficient, which determine the degree of impact of the regressors to the dependent variable and $\boldsymbol{\varepsilon}$ represents the error term.

The aim of the analysis is to identify, with the information at disposal in the database, the relationship – if any – between the technical specifications of an infrastructure, the environmental and exogenous conditions where the infrastructure is deployed, its cost.

With a sufficiently detailed number of observations for all the regressors, it is therefore aimed to determine the cost of the infrastructure.

1.3. Linear regression analysis

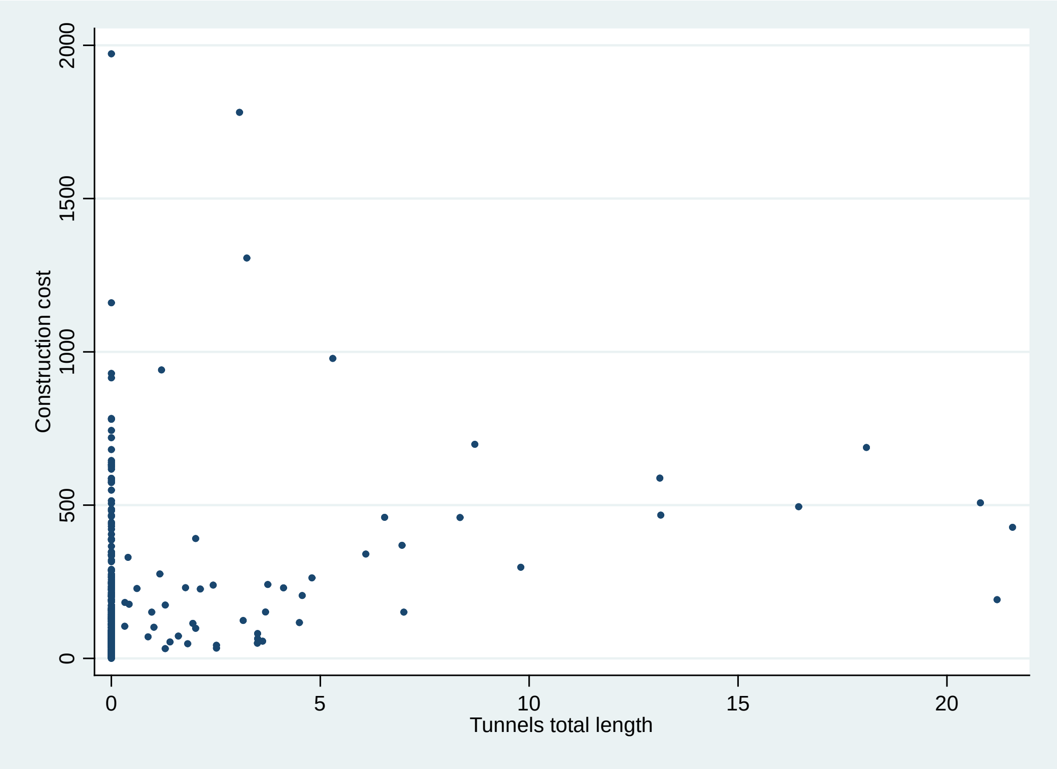
The linear regression analysis is performed considering the whole set of data available, to identify a universal formula that explains the relationship between the cost of a line, its length and the factors that are supposed to make the cost vary.

The factors are selected to be both relative to:

- the features of the infrastructure: i.e. length, length of tunnels, length of bridges, length of viaducts, number of interfaces, number of stations, infrastructure type, work type.
- the conditions of the environment: i.e. urban/ rural territory; flat/ hilly/ mountainous terrain.

The lack of information on the data, requires assumptions to be made. In case of missing information on the presence of structures (i.e. bridges, viaducts, etc.), it is necessary to decide how to treat the data. Would the data be considered to be missing because of no structures on the project, all those cases would equal zero. As a result, the regression analysis may result biased as a significant concentration of values would occur in the range of zero structures. A graphical example of the case is reported in figure below, where the observations are seen in terms of cost and presence of tunnels. Considering as equal to zero the length of tunnels from observations not providing any length value presents a disproportionate concentration of observation on the vertical line relative to $x=0$, which is not realistic.

Figure 1 - Construction cost and tunnel length

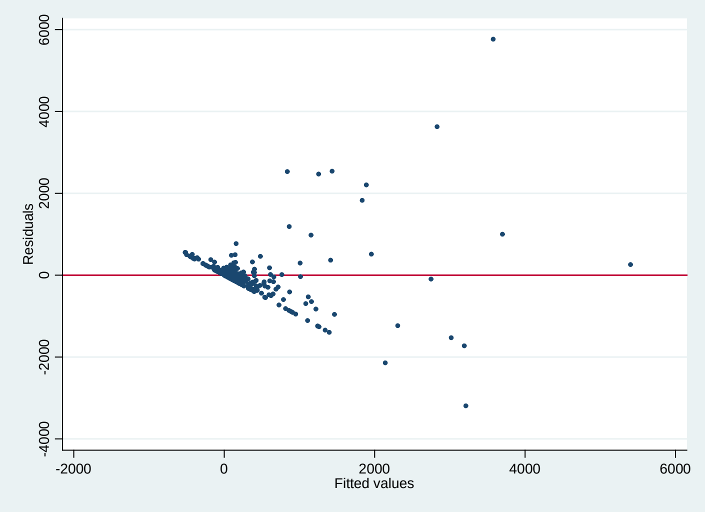


The results from regressing all the elements together without performing any cleaning of the observations does not present consistent result (Exp_1).

To avoid the bias, the alternative is represented by considering the missing data as resulting from poor data quality and eventually drop the entire observation. The approach eliminates the bias, but it also significantly reduces the amount of data, which are used to analyse the costs (Exp_2).

While the level of information reduces, the quality of the results increases. Nonetheless, the regression still lacks of significance when considering the presence of heteroskedasticity, which compromises the reliability of the analysis.

Figure 2 - Plot of the residuals



Would it not have been the case, it would still have made little sense to consider such a broad range of different type of works included in a single regression analysis without discriminating among them.

To attempt the discrimination, it can be considered to substitute the total length with the breakdown of lengths of the different types of infrastructure work: new lines, rehabilitation and upgrade, signalling, electrification (Exp_3) and further to differentiate high-speed and conventional lines (Exp_4).

	Exp_1	Exp_2	Exp_3	Exp_4
Line length	0.20**	2.13***		
New_line			6.45***	
New_line_hs				6.63***
New_line_conv				0.73
Rehab_km			-0.38	-0.28
Signalling_km			0.0077	-0.02
Electrification_km			2.20***	2.14***
Tunnel_length	44.17***	40.16***	26.17***	25.89***
Bridges&viaducts_length	13.80***	10.92**	0.95	0.31
N_interfaces	0.19	-3.78	0.12	0.12
N_stations	1.24	-5.40***	0.81	0.82
Urban (dummy)	-344.55***	-344.36***	-272.51***	-290.34***
Rural (dummy)	21.22	20.27	-207.54**	-204.37***
Terrain_flat (dummy)	336.86***	321.84***	323.12***	345.02***
Terrain_hilly (dummy)	-113.44	-87.72	23.69	34.01
Terrain_mount (dummy)	334.94***	351.99***	272.72**	260.22**
Categ_hub (dummy)	44.61	126.55	133.52	133.00
Categ_bridge (dummy)	-27.23	67.74	198.7	169.06
Categ_bypass (dummy)	27.09	115.48	84.86	91.00
Categ_Tunnel (dummy)	5.91	73.16	353.74***	360.2***
_const	83.60**	8.18	-4.68	12.74
N	498	294	294	294
R^2	0.4171	0.4775	0.6979	0.7027
Heteroskedasticity	Y	Y	Y	Y

A very general approach that does not differentiate sufficiently the types of infrastructure would not lead to any significant result (Exp_1 and Exp_2). This is due from very different cases being treated as if they were comparable. As a result, breaking down the variable (in the case, the length of the infrastructure) into sub-components made it possible to identify how the observation relative to new high-speed lines are very different from e.g. signalling ones.

To be said, all the regressions show significant heteroskedasticity. OLS analysis is not optimal when heteroskedasticity is present, as OLS analysis gives equal weight to all observations when, in fact, observations with larger disturbance variance contain less information than observations with smaller disturbance variance.[1]

The overall, general analysis suffers from the aim to be very broad. It indeed includes in the same analysis very different types of infrastructure and work. Breaking down the regression into pieces led to generate more realistic results on specific cases (i.e. high-speed new lines construction), but provides very inconsistent results on many variables (e.g. the dummies on terrain, the categories, etc.) and is affected by very high heteroskedasticity.

To the aim of the analysis, the approach is thus to break down the clusters on which the analyses are carried out and concentrate on isolating the correlation between the variables comparing similar observations (i.e. new high-speed lines with new high-speed lines; electrification works with electrification works; and so on) (see Second Progress Report).

[1] Paul Allison, Professor of Sociology at the University of Pennsylvania.