Quantitative Methods - Coursework 1

Yafei Ye, MRes Spatial Data Science and Visualisation, CASA, UCL

How do different types of budgets allocated to the road safety initiative influence the death and serious injury on the road?

Introduction

How different parts of road safety initiative budgets influence the change rate of the counts of those killed and seriously injured on the road is the research question of this coursework. Part one explains how to deal with the data to make it suitable for analysis; part two introduces the analysis procedure; part three shows the result; part four makes the conclusion.

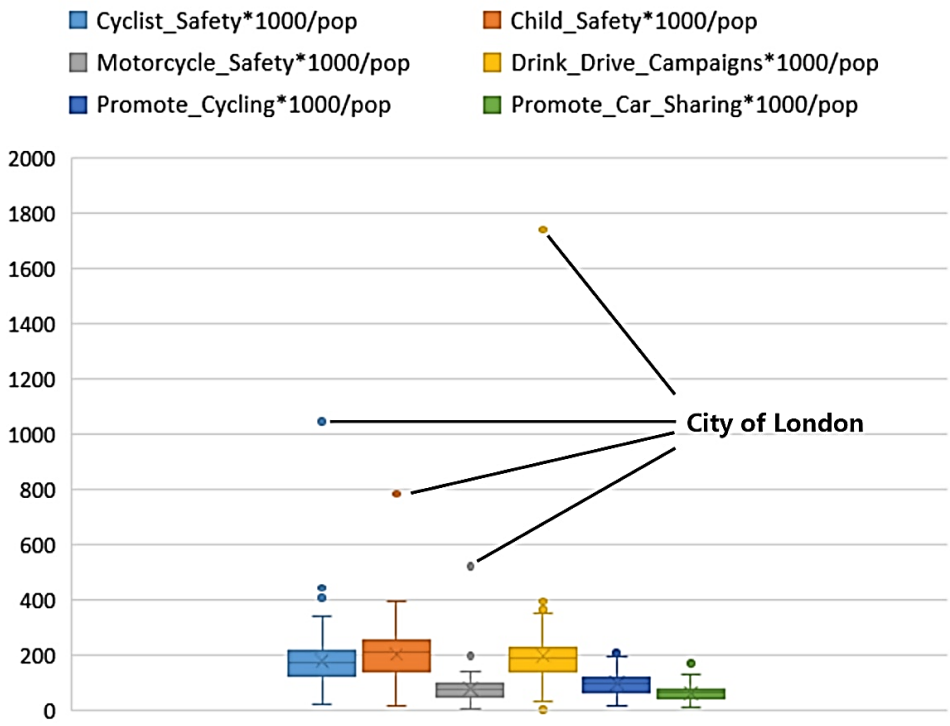
1 Data

According to the data provided, the budgets allocated to the road safety initiative in local authority areas in 2009 are divided into 6 parts. To fit the time of the budgets, *2009_KSI* and *2008_KSI* are chosen to calculate the change rate. To eliminate the impact of population, both the budget and the KSI are divided by the population. The formula of the change rate shows like below:

*Change Rate = [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000*

Through making Box and Whisker, it finds that the budgets of City of London are outliners in 4 subcategories which may make huge impact for final results. Therefore, City of London is removed.

Figure 1 – Box and Whisker for 6 subcategories of budgets



The descriptive statistics of final data are shown as below:

Figure 2 – Descriptive Statistics of Final Data

	N	Minimum	Maximum	Mean	Std. Deviation
[(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000	150	-41.28%	6.89%	-5.2412%	7.80789%
Cyclist_Safety*1000/pop	150	22.33389	442.70833	173.41476	71.72078
Child_Safety*1000/pop	150	16.47446	396.68905	199.28992	76.66523
Motorcycle_Safety*1000/pop	150	4.442470	208.33333	75.632871	34.13137
Drink_Drive_Campaigns*1000/pop	150	4.685304	393.95005	186.02377	66.80006
Promote_Cycling*1000/pop	150	16.73963	230.59867	97.221341	43.38007
Promote_Car_Sharing*1000/pop	150	10.42753	171.33956	62.722295	26.16686
Valid N (listwise)	150				

2 Analysis

2.1 Constructing Multiple Regression Model

To analyze how different budgets influence the change rate of KSI, we set the change rate as the Dependent Variable, the 6 parts of budgets as the Predictors, to make multiple regression. The modelled relationship are shown as below:

$$KCR = \beta_0 + \beta_1 * CYS + \beta_2 * CHS + \beta_3 * MS + \beta_4 * DDC + \beta_5 * PC + \beta_6 * PCS + \varepsilon^1$$

The result of Adjusted R Square shows that this model can explain 7.2% of the variability in KSI change rate (Figure 3).

Figure 3 – Model 1 Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.330[a]	.109	.072	7.52332%	1.747

a. Predictors: (Constant), Promote_Car_Sharing*1000/pop, Motorcycle_Safety*1000/pop, Drink_Drive_Campaigns*1000/pop, Promote_Cycling*1000/pop, Cyclist_Safety*1000/pop, Child_Safety*1000/pop

b. Dependent Variable: [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000

[1] KCR = [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000; CYS = Cyclist_Safety*1000/pop; CHS = Child_Safety*1000/pop; MS = Motorcycle_Safety*1000/pop; DDC = Drink_Drive_Campaigns*1000/pop; PC = Promote_Cycling*1000/pop; PCS = Promote_Car_Sharing*1000/pop

2.1.1 Durbin-Watson Test

H_0: No first order autocorrelation

H_1: First order correlation exists , $\alpha = 0.05$

Looking up the table to know[2], 1.651 < Durbin-Watson value = 1.747 (Figure 3) < 1.817, pass the test and there is not enough evidence to reject H_0. Therefore, there is no first order autocorrelation in this regression in 0.05 significance.

Figure 4 – Model 1 ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	989.660	6	164.943	2.914	.010
	Residual	8093.843	143	56.600		
	Total	9083.502	149			

a. Dependent Variable: [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000

b. Predictors: (Constant), Promote_Car_Sharing*1000/pop, Motorcycle_Safety*1000/pop, Drink_Drive_Campaigns*1000/pop, Promote_Cycling*1000/pop, Cyclist_Safety*1000/pop, Child_Safety*1000/pop

2.1.2 F-Test

H_0: All coefficients = 0

H_1: At least one $\beta j \neq 0$, $\alpha = 0.05$

F Sig. = 0.010 (Figure 4) < α, reject H_0, accept H_1。 Therefore, the independent variables in whole are linearly correlated with the dependent variable at a significant level of 0.05.

Figure 5 – Model 1 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	7.082	5.962		1.188	.237					
	Cyclist_Safety*1000/pop	-.029	.010	-.266	-2.984	.003	-.202	-.242	-.24	.786	1.272
	Child_Safety*1000/pop	-.021	.010	-.205	-2.033	.044	-.064	-.168	-.16	.615	1.626
	Motorcycle_Safety*1000/pop	-.014	.018	-.062	-.782	.436	-.048	-.065	-.06	.977	1.024
	Drink_Drive_Campaigns*1000/pop	-.026	.011	-.219	-2.299	.023	-.079	-.189	-.18	.688	1.454
	Promote_Cycling*1000/pop	.013	.015	.072	.855	.394	.168	.071	.067	.881	1.135
	Promote_Car_Sharing*1000/pop	.023	.024	.076	.941	.348	.127	.078	.074	.951	1.052

a. Dependent Variable: [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000

2.1.3 T-Test

H_0: Gradient = 0

[2] http://web.stanford.edu/~clint/bench/dw05b.htm

H_1: Gradient $\neq$ 0 , $\alpha = 0.05$

CYS p-value = 0.003 < α, CHS p-value = 0.044 < α, MS p-value = 0.436 > α, DDC p-value = 0.023 < α, PC p-value = 0.394 > α, PCS p-value = 0.348 > α (Figure 5). For CYS, CHS and DDC, reject H_0, accept H_1。 For MS, PC and PCS, there is not enough evidence to reject H_0. Therefore, MS, PC and PCS have no significant linear relationship with the dependent variable. We consider to eliminate them to construct a new model later.

Figure 6 – Correlations

		[(2009_KSI /2009_pop)/ (2008_KSI/ 2008_pop) -1]*1000	Cyclist_Safety *1000/pop	Child_Safety *1000/pop	Motorcycle_ Safety*1000/ pop	Drink_Drive _Campaigns *1000/pop	Promote_Cycl ing*1000/pop	Promote_Car_ Sharing*1000/ pop
[(2009_KSI/200 9_pop)/ (2008_KSI/2008 _pop)-1]*1000	Pearson Correlation	1	-.202*	-.064	-.048	-.079	.168*	.127
	Sig. (2-tailed)		.013	.435	.560	.336	.039	.121
	N	150	150	150	150	150	150	150
Cyclist_Safety*1 000/pop	Pearson Correlation	-.202*	1	-.222**	-.010	-.165*	-.169*	-.085
	Sig. (2-tailed)	.013		.006	.907	.044	.038	.301
	N	150	150	150	150	150	150	150
Child_Safety*10 00/pop	Pearson Correlation	-.064	-.222**	1	-.102	-.456**	-.172*	-.164*
	Sig. (2-tailed)	.435	.006		.216	.000	.036	.045
	N	150	150	150	150	150	150	150
Motorcycle_Safe ty*1000/pop	Pearson Correlation	-.048	-.010	-.102	1	.017	-.056	-.014
	Sig. (2-tailed)	.560	.907	.216		.834	.495	.863
	N	150	150	150	150	150	150	150
Drink_Drive_Ca mpaigns*1000/p op	Pearson Correlation	-.079	-.165*	-.456**	.017	1	-.018	.066
	Sig. (2-tailed)	.336	.044	.000	.834		.826	.420
	N	150	150	150	150	150	150	150
Promote_Cycling *1000/pop	Pearson Correlation	.168*	-.169*	-.172*	-.056	-.018	1	.118
	Sig. (2-tailed)	.039	.038	.036	.495	.826		.152
	N	150	150	150	150	150	150	150
Promote_Car_Sh aring*1000/pop	Pearson Correlation	.127	-.085	-.164*	-.014	.066	.118	1
	Sig. (2-tailed)	.121	.301	.045	.863	.420	.152	
	N	150	150	150	150	150	150	150

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

2.1.4 Correlations

The result (Figure 6) shows CYS and CHS, CHS and DDC are significantly correlated at the 0.01 level. However, none of Pearson Correlation values are larger than 50%. All VIF values in Figure 5 are also smaller than 10, which proves that there in no serious multicollinearity problem. As a result, we reserve CYS, CHS and DDC to try to build a better model.

2.2 Using Benford's Law to identify artificial data

The coefficients of PC and PCS in Figure 5 are positive, which means along with the increase of the budgets in cycling promotion and car sharing promotion, the KSI will also increase to some extent, which doesn't fit well with common sense. Therefore, we use Benford's Law to identify whether they are artificial data.

2.2.1 Chi-Squared Test

H_0: First digit distribution of the budget is drawn from a binomial distribution with Benford's Law
H_1: First digit distribution of the budget is not drawn from a binomial distribution with Benford's Law , $\alpha = 0.05$
PC Chi-Square p-value = $2.36*10^{-8} < \alpha$, PCS Chi-Square p-value = $1*10^{-13} < \alpha$ (Figure 7a, 8a). Therefore, reject H_0, accept H_1. Both the first digit distribution of PC and PCS are not following the Benford's Law well. They may be artificial data.

Figure 7a – First Digit Distribution between Budget Allocated to Cycling Promotion Schemes and Benford's Law

Budget Allocated to Cycling Promotion Schemes (£)				
	First Digit		Benford's Law	
Digit	Actual	Actual %	Expected	Expected %
1	69	45.70%	45.451	30.10%
2	6	3.97%	26.5911	17.61%
3	7	4.64%	18.8599	12.49%
4	10	6.62%	14.6319	9.69%
5	10	6.62%	11.9592	7.92%
6	12	7.95%	10.1019	6.69%
7	10	6.62%	8.758	5.80%
8	11	7.28%	7.7312	5.12%
9	16	10.60%	6.9158	4.58%
Sum	151	100.00%	151	100.00%
Chi-Square p-value			0.0000000236	

Figure 7b - First Digit Distribution Fequency between Budget Allocated to Cycling Promotion Schemes and Benford's Law

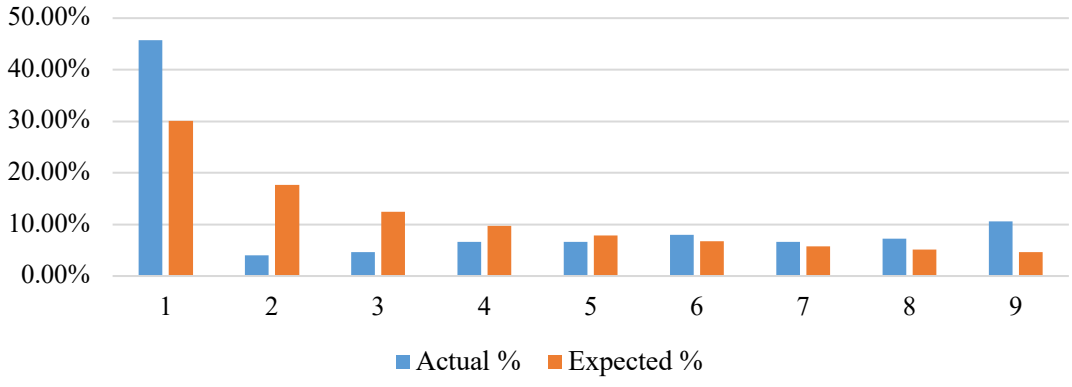
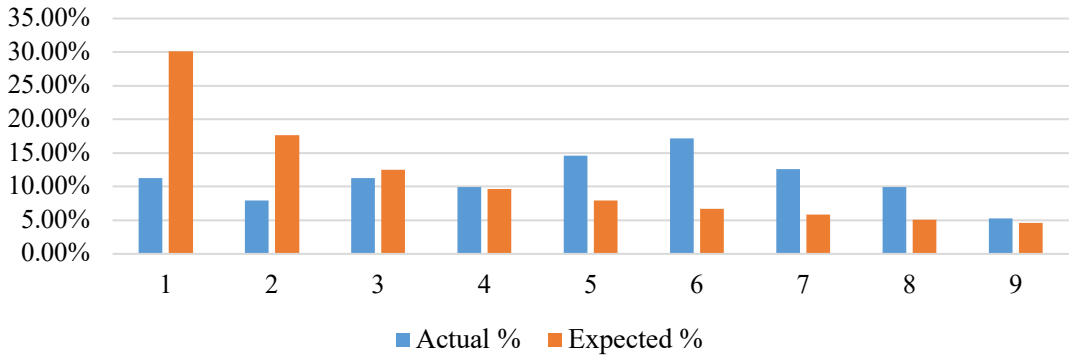


Figure 8a – First Digit Distribution between Budget Allocated to Car Sharing Promotion and Benford's Law

Budget Allocated to Car Sharing Promotion schemes (£)				
First Digit			Benford's Law	
Digit	Actual	Actual %	Expected	Expected %
1	17	11.26%	45.451	30.10%
2	12	7.95%	26.5911	17.61%
3	17	11.26%	18.8599	12.49%
4	15	9.93%	14.6319	9.69%
5	22	14.57%	11.9592	7.92%
6	26	17.22%	10.1019	6.69%
7	19	12.58%	8.758	5.80%
8	15	9.93%	7.7312	5.12%
9	8	5.30%	6.9158	4.58%
Sum	151	100.00%	151	100.00%
Chi-Square p-value			0.00000000000010	

Figure 8b - First Digit Distribution Fequency between Budget Allocated to Car Sharing Promotion Schemes and Benford's Law



2.3 Reconstructing Multiple Regression Model

We set the change rate as the Dependent Variable, the CYS, CHS and DDC as the Predictors, to make multiple regression. The modelled relationship are shown as below:

$$KCR = \beta_0 + \beta_1 * CYS + \beta_2 * CHS + \beta_3 * DDC + \varepsilon$$

The result of Adjusted R Square shows that this model can explain 7.5% of the variability of KCR (Figure 9) which is higher than Model 1. The same as the tests in Model 1, this model can pass both DW Test (Figure 9) and F-test (Figure 10), and all predictors can pass T-test and VIF Test (Figure 11). The Residual Distribution Frequency Histogram (Figure 11) is normally distributed, which proves the regression result is reliable.

Figure 9 – Model 2 Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
2	.305[a]	.093	.075	7.51109%	1.724

a. Predictors: (Constant), Drink_Drive_Campaigns*1000/pop, Cyclist_Safety*1000/pop, Child_Safety*1000/pop

b. Dependent Variable: [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000

Figure 10 – Model 2 ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	846.695	3	282.232	5.003	.002[b]
	Residual	8236.807	146	56.416		
	Total	9083.502	149			

a. Dependent Variable: [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000

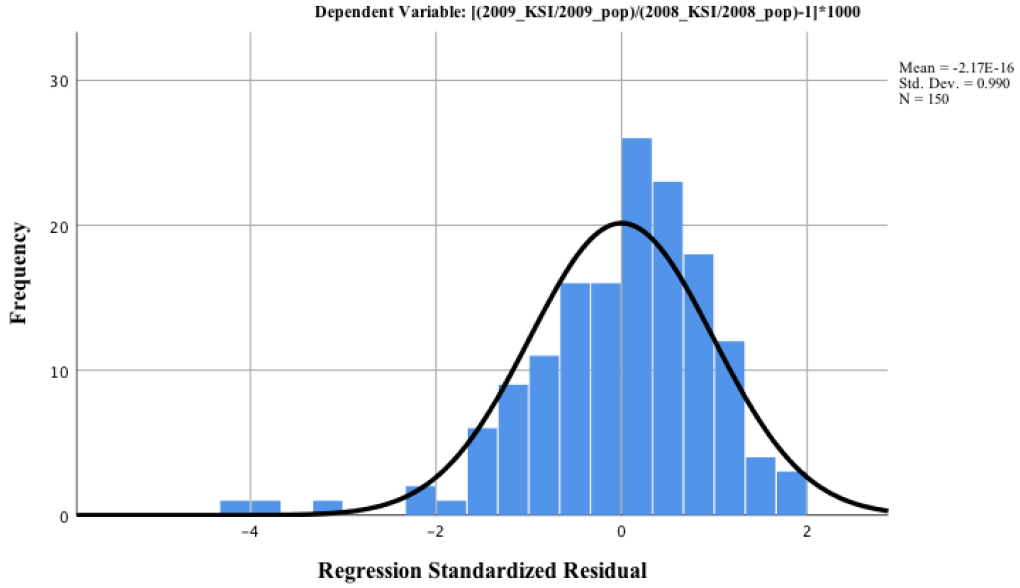
b. Predictors: (Constant), Drink_Drive_Campaigns*1000/pop, Cyclist_Safety*1000/pop, Child_Safety*1000/pop

Figure 11 – Model 2 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
2	(Constant)	10.233	4.326		2.365	.019					
	Cyclist_Safety *1000/pop	-.032	.009	-.294	-3.458	.001	-.202	-.275	-.273	.861	1.161
	Child_Safety* 1000/pop	-.024	.010	-.237	-2.517	.013	-.064	-.204	-.198	.701	1.427
	Drink_Drive_ Campaigns*10 00/pop	-.028	.011	-.236	-2.532	.012	-.079	-.205	-.200	.717	1.394

a. Dependent Variable: [(2009_KSI/2009_pop)/(2008_KSI/2008_pop)-1]*1000

Figure 12 – Residual Distribution Frequency Histogram



3 Result

Finally, the multiple linear regression equation of Model 2 is established:

$$KCR = 10.233 - 0.032*CYS - 0.024\beta_2*CHS - 0.028*DDC + \varepsilon$$

4 Conclusion

CYS, CHS and DDC can influence KCR negatively, and all of them can explain 7.5% of the variability of KCR. The influence of CYS is a little more significant than CHS and DDC as the absolute value of its coefficient is a bit larger, while MS, PC and PCS have no significant influence, and PC and PCS may be artificial data.