Candidate 2 evidence

Word count: 881

Introduction

This project is on road safety in England, Wales and Scotland. The data is taken from a data set called Road safety data on "**data.gov.uk**" it uses the "STATS 19" accident reporting form. The statistics relate to accidents on public roads that are reported to the police. It only applies to accidents where personal injury is received. The data set starts from January 1st 1979, and ends December 31st 2015 (inclusive).
The data will be tested to see if there is enough evidence to suggest a linear relationship, through linear regression analysis, between time in half decades, and severity; where a severity of 1 is a fatality, 2 is a serious casualty, and 3 is a slight casualty. A linear relationship is expected. The data will also be tested to see if there is a significant improvement in severity as expected, between 1981 and 2015. Only the data from 1981-2015 will be used here. The data will first be represented by a scatter plot, to judge by eye if there is a linear relationship between the two variables described above, before calculating statistical information. The severity will be used, as it best describes the likelihood of a fatality on the roads rather than the number of injuries that occurred. This, when looked at in total number, could describe a large amount of slight injuries, but a smaller total number of fatal injuries, while another year could have the majority of those injuries being serious, with a smaller number of casualties. The severity used for each year will be an average of all the severities listed in a year, and the average severity for each half decade will be found the same way, but by every five years.

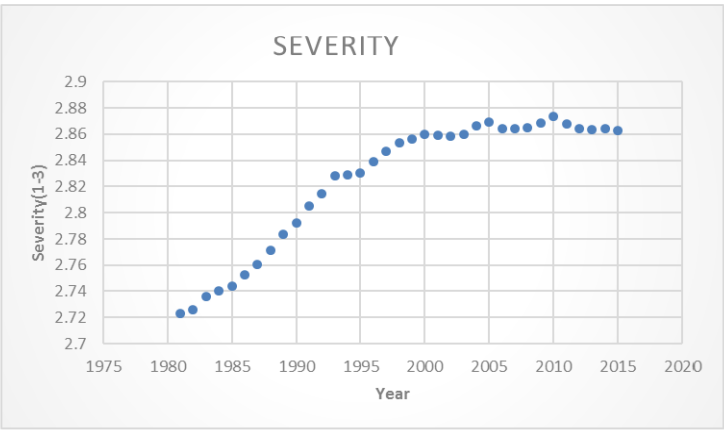
Sampling

The data from data.gov.uk is split between many files, and some of these files have too many rows for excel to be opened. This required a great deal of organizing, until the data was sorted by year. There was a need for only one of the columns from the data, so the others were deleted - *the organized data for this specific topic can be found linked in the references.*

Once the data had been separated by year, the data was then arranged again, but now into half decades (5 full years), starting from 1981-1985. This was done to simplify the tests done on the data to make it more practical. This sampling method smooths the distribution, but is still a relatively accurate representation of the relationship between time and severity, as can be seen by the graphs below.

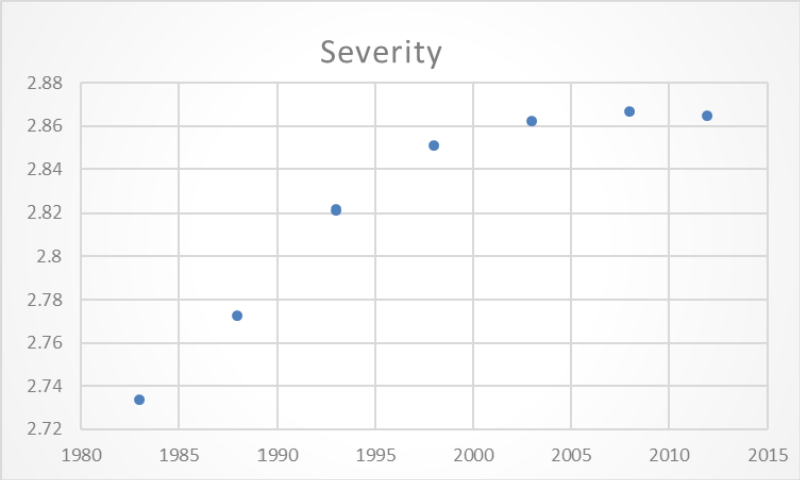
Presenting and analysing data

Scatterplot of Severity against time by year.



SEVERITY

- 1 = Fatal
- 2 = Serious
- 3 = Slight



This is a representation of the graph above where the 35 years have been grouped into seven groups of 5 full years in length.

The second graph clearly has a very similar distribution, but the curve has been smoothed.

There does not appear to be a linear relationship between severity and year, between 1981 and 2015, but there seems to be a section of the graph between 1981 and 1998 where it could be linear. This can be seen above in graph 2. Though the whole graph will be tested to make 100% sure.

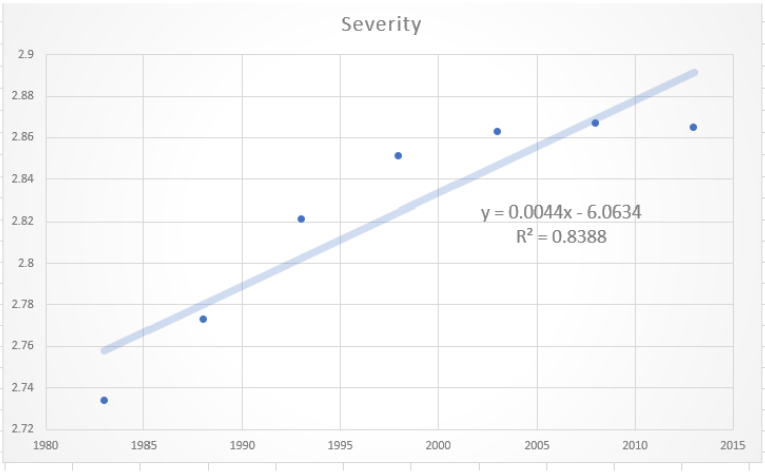
Data—

HALF DECADES	SEVERITY
1981-1985	2.733692
1986-1990	2.772621
1991-1995	2.821315
1996-2000	2.851012
2001-2005	2.862519
2006-2010	2.867035
2011-2015	2.864612

CALCULATED VALUES—

The equation of the linear regression line that could describe the data was calculated and drawn on the whole graph below.

$$y = 0.0044x - 6.0634$$



The plotted line appears to emphasize a pattern in the data, but the calculated r value as an indication of the strength of the relationship between year and accident severity is strong.

The r value was also obtained from $R^2 = 0.8388$. This gives an r value of 0.92. This suggests that the model does have a strong, positive linear correlation (since the value is relatively close to positive one) between the year and severity of accident. As the year increases the number associated with the severity of accident increases.

Conclusion

In conclusion, there is a linear relationship between time and severity, as seen by the scatterplot, and the value of r. There is enough evidence to suggest a linear relationship between the time in half decades and accident severity; where a severity 1 is fatality, 2 is a serious casualty, and 3 is a slight casualty. There is a significant difference in the severity of an average casualty in 1981 and 2015. This suggests that road safety from 1981 to 2015 has increased, as the chance of an accident being fatal has significantly decreased since 1981.