Statistical Report

Lu
Tingjia
5306

Problem

Fire damage in the United States amounts to billions of dollars, much of it insured. The time taken to arrive at the fire is critical. This raises the question, should insurance companies lower premiums if the home to be insured is close to a fire station? To help make a decision, a study was undertaken wherein a number of fires were investigated. The distance to the nearest fire station (in miles) and the percentage of fire damage were recorded. We would like to leverage the data to answer this question: Should insurance companies lower premiums if the home to be insured is close to a fire station?

Executive summary

This report summarizes my findings on potential association between distance to the nearest fire station and the percentage of fire damage. The report includes the following: 1) a brief explanation of the study, 2) the data, 3) statistical methods used, 4) some assumptions used, and 5) a summary of result.

The study

The first step is to transform the question into a statistical problem which is whether the percentage of fire damage will be smaller when the distance to the nearest fire station is smaller. If the percentage of fire damage will be smaller when the distance to the nearest fire station is smaller which means the distance and the percentage are positively related so that the insurance companies should lower premiums if the home to be insured is close to a fire station. In this report 85 samples are used to answer this question. Firstly, I draw the scatter plot of the 2 variables and then I calculate the coefficient of correlation of the 2 variables and the regression formula to quantity the linear association of the 2 variables.

The data

The data collected the distance to the nearest fire station (in miles) and the percentage of fire damage of 85 in the sample. Exhibit 1 presents data of the variable Distance and Percent. Summary statistics for the variable Distance and Percent are given in Exhibit 2. From Minimum, median,3rd quantile and maximum value of variable Distance and Percent we can see there might be a positive relation between variable Distance and Percent but we can't use this as an evidence and the strong evidence should refer to next part" Statistical methodology". In next part we use variable Distance as predictor variable and variable Percent as response variable.

Statistical methodology

My main objective is to verify whether variable Distance and variable Percent has relationship and what kind of relationship they have, in further I would like to quantity the relationship. In this course we discussed how to quantity the association between two quantitative variables. Usually we tend to draw a scatter plot to see the direction, form and strength of the association between 2 variables. Figure 1 shows a scatterplot of the data where the x-coordinate represents the distance to the nearest fire station and y-coordinate gives the percentage of the fire damage. The least-squares regression line is also graphed on Figure 1. The Scatterplot suggests a positive linear relationship: large distance to the nearest fire station(in
miles) with large percentage of damage & small distance to the nearest fire station (in miles) with small percentage of damage. But the relationship is not perfect. There is more variability in percentage of damage when distance to the

nearest fire station (in miles) is between 4 and8. Exhibit 3 shows the sample coefficient of correlation between Distance and Percent is 0.707which shows a strong positive linear association. To supplement this graphical analysis, I give the simple linear regression formula and the hypothetical test of parameters in Exhibit 4. The Regression model is:

$$\text{Percent} = \beta 0 * \text{Distance} + \beta 1,$$
$$\text{Percent} = 5.35 * \text{Distance} + 23.1,$$

And we should reject the null hypothesis of $\beta 0 = 0$, $\beta 1 = 0$ under significance level 0.05, which further shows the strong positive linear association between distance to the nearest fire station(in miles) and percentage of fire damage.The necessary conditions to apply these methods above are verified next.

Assumptions

To obtain the correlation coefficient and check that method assumptions are met, R was used (see appendix for more details about R code). First we discuss the assumptions. Above, we don't remove any sample point and no changes for the original data which means we assume there is no noise in the data. On the other hand, it is reasonable to assume that the assessment of any entry in the data is independent of the assessment of the others and here we don't take other factors that may result in different fire damage into consideration. Randomization Condition is satisfied, our 85 samples are a random sample from the population of interest.

Finally, the assumption of Normality was checked graphically with the aid of Q-Qplot andBoxplot: see Figures 2,3 respectively. All graphs suggest that the Normality assumption is reasonable for the regression of distance to the nearest fire station (in miles) and percentage of fire damage. Boxplot confirms the symmetry of response variable and doesn't identify any outliers. The Q-Q plot doesn't provide evidence of an important departure fromNormalityassumption.

Results

The output provides evidence that small distance to the nearest fire station (in miles) with small percentage of damage. The correlation coefficient between distance to the nearest fire station (in miles) and percentage of fire damage is 0.707 and the linear regression formula represents the percent of damage will increase 5.35 when the distance to the nearest fire station increases 1 mile. We also conduct the hypothetical test of the regression parameter and result shows the two parameters can't be zero under significance level 0.05. Now we can answer the original question that insurance companies should lower premiums if the home to be insured is close to a fire station because it is about to have a smaller percentage of fire damage.

Appendix

Exhibit 1

```
> Distance
 [1] 7.5 8.3 6.2 1.6 5.6 6.0 4.3 8.1 5.7 0.3 1.6 2.5 5.8 5.3 6.3 3.4
[17] 6.2 3.2 6.3 6.1 4.6 6.7 0.5 3.2 5.3 5.0 4.8 6.4 0.2 4.9 6.2 7.6
[33] 6.6 2.9 2.1 4.8 8.1 1.2 4.6 4.0 6.1 0.8 5.9 6.5 6.5 7.5 7.2 6.7
[49] 4.1 4.0 4.8 4.0 6.2 5.5 7.0 7.5 6.2 5.7 7.3 1.9 2.8 6.3 9.4 3.7
[65] 4.9 1.8 3.6 2.9 2.6 3.1 4.7 5.9 2.5 4.6 5.2 7.6 3.7 7.4 3.3 5.1
[81] 5.1 7.5 4.7 2.7 2.7
> Percent
 [1] 68 66 34 30 70 62 47 72 40 53 18 48 53 48 64 52 61 34 65 66 33
[22] 76 34 46 55 33 46 49 17 47 63 69 54 52 43 35 58  5 46 57 40 39
[43] 42 62 52 76 67 45 23 33 59 50 49 44 52 68 48 55 77  9 35 51 84
[64] 30 61 40 24 37 51 30 47 54 47 52 52 52 33 74 56 53 54 76 37 45
[85] 50
```

Exhibit 2

```
> summary(Distance)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.200   3.300   5.100   4.885   6.300   9.400
> summary(Percent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   40.00   50.00   49.22   59.00   84.00
```

Exhibit 3

```
> cor(Distance,Percent)
[1] 0.7073615
```

Exhibit 4

```
> summary(mod)

Call:
lm(formula = Percent ~ Distance)

Residuals:
    Min      1Q  Median      3Q     Max
-24.523  -8.415   1.091   8.220  28.290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.1065     3.1076   7.435 8.56e-11 ***
Distance      5.3467     0.5865   9.117 3.81e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.11 on 83 degrees of freedom
Multiple R-squared:  0.5004,    Adjusted R-squared:  0.4943
F-statistic: 83.12 on 1 and 83 DF,  p-value: 3.808e-14
```

Figure 1: Scatterplot (Distance vs Percent)

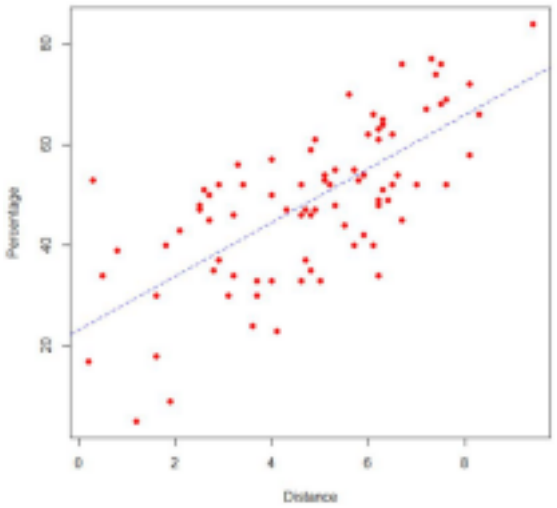

Figure 2: QQ plot of residual values

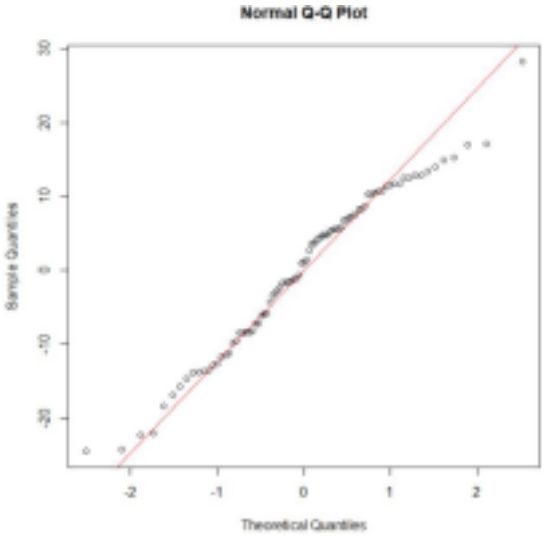

Figure 3: Boxplot of the residual values