

## **CASE STUDY**

### **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.

## **Executive summary**

This report summarizes our findings on whether or not there is a correlation between the distance to the nearest fire station and the percentage of fire damage. The report consists of the following: 1) a brief description of the study, 2) a presentation of the data, 3) the statistical methodology employed to conduct the data analysis, 4) the assumptions made about the given data set, and 5) a summary of our findings.

## **The study**

Insurance companies offer home insurance coverage, according to which they must reimburse the financial losses incurred from damages to properties. When calculating premiums, insurance companies take numerous factors into consideration and the risks associated with them. Given that, during a fire incident, damages to a property increase with time, we hypothesize that the percentage of fire damage to properties will be influenced by the proximity of the properties to fire stations. To test this, a random sample of 85 fire incidents was investigated, from which the distance to the nearest fire station and the percentage of fire damage were recorded. We then used the collected data to perform correlation analysis to conclude whether we can establish any relationship between the distance to the nearest fire station and the percentage of fire damage. Additionally, we used hypothesis testing to determine whether the relationship in the sample data effectively models the relationship in the population.

## **The data**

The data collected from the investigation of fire accidents includes 85 observations, where each observation provides us with the distance to the nearest fire station and the percentage of fire damage recorded. Exhibit 1 presents the aforementioned data. Exhibit 2 presents the summary statistics for each of the variables in our data set. However, this is not enough for us to establish any relationship between the two variables. For that reason, we perform a correlational analysis of our data in the next section. Furthermore, we are only provided with sample data, meaning if we were given a different sample, we would obtain a different correlational analysis and potentially different conclusions. And since we want to draw conclusions about the entire population, we will conduct a hypothesis test to decide whether the relationship in the sample data is strong enough to infer a relationship in the population. This is also presented in the next section.

## **Statistical methodology**

The main objective of this study is to establish whether there is any statistical relationship between the two variables: the distance to the nearest fire station (explanatory variable) and the percentage of fire damage recorded (response variable). Figure 1 shows a scatterplot of our variables against each other, from which it is evident that there is a linear positive correlation between them. Figure 1 also plots a regression line using the least-squares method. Since the

association between the variables in the sample data is linear, we can compute the measure of the strength of this association using Pearson's correlation coefficient (denoted by  $r$ ). Note that the correlation coefficient is sensitive and is affected by extreme outliers; however, since there are no apparent outliers in Figure 1, we do not have to adjust our data set to remove any outliers.

The sample correlation coefficient,  $r$ , is an estimation of our population correlation coefficient, conventionally denoted by  $\rho$ . To check the effectiveness of  $r$  in estimating  $\rho$ , we conduct a hypothesis test using our sample correlation coefficient and our sample size  $n$ , where our null hypothesis states that the population correlation coefficient  $\rho$  is not significantly different from 0, and our alternative hypothesis states that the population correlation coefficient  $\rho$  is significantly different from 0.

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

Using the  $p$ -value method (utilizing  $t$ -distribution), if our  $p$ -value is less than our significance level, then we reject our null hypothesis and accept the alternative hypothesis; otherwise, we do not reject our null hypothesis.

### Assumptions

The inference-making technique we present in our study (hypothesis testing) is valid only under the following assumptions about random errors of our simple linear regression:

- a) Random errors approximately follow a normal distribution
- b) Random errors have mean zero
- c) Random errors have a constant variance
- d) Random errors are uncorrelated/don't follow any pattern.

These assumptions about random errors can be checked using residual plots (Exhibit 3).

Assumption a) can be checked graphically using a combination of a histogram, a boxplot, and a Q-Q plot, as shown in Figures 2, 3, and 4, respectively. Assumptions b) and c) can be checked graphically using a plot of residuals vs. the explanatory variable (distance to the nearest fire station), as shown in Figure 5, where it is evident that the residual points are plotted randomly around the zero line, as well as evenly spread out around the zero line. Additionally, the assumption b) can be checked using the summary statistics for the residual values (Exhibit 3), where it states that the mean value is zero. Lastly, assumption d) can be checked graphically using Figure 5, where it is evident that the residual values don't follow any specific pattern/are uncorrelated.

Since all the required assumptions are met, we proceed to compute the sample correlation coefficient and to carry out the hypothesis test. The results and conclusions are discussed in the next section.

## Results and Future work

Exhibit 4 shows our calculations for computing the sample correlation coefficient and for carrying out the hypothesis test. From the output, we can see that our sample correlation coefficient is 0.7109, implying that there is a strong correlation between the distance to the nearest fire station and the percentage of fire damage.




For our hypothesis testing, we use a significance level of  $\alpha = 0.05$ . From the output, we can also see that  $p\text{-value} = 1.2495 \cdot 10^{-14} \leq \alpha$ , from which it follows that we can reject our null hypothesis and accept the alternative hypothesis. In other words, we have sufficient evidence to conclude that there is a significant linear relationship between the distance to the nearest fire station and the percentage of fire damage because the correlation coefficient is significantly different from zero.

Now, we can answer the posed question “Should insurance companies lower premiums if the home to be insured is close to a fire station?”. From the obtained results, insurance companies should indeed lower premiums for homes that are close to a fire station.

This study could be further extended by finding the equation for the simple linear regression and using the said equation to calculate the amount by which the insurance premiums should be lowered depending on the distance from the nearest fire station. Additionally, we could look into other possible factors that could impact the percentage of fire damage, and extend the linear regression into multiple linear regression, where there are two or more explanatory/independent variables.

## Appendix

### Exhibit 1:

	Distance 	Percent 						
1	7.5	68	31	6.2	63	61	2.8	35
2	8.3	66	32	7.6	69	62	6.3	51
3	6.2	34	33	6.6	54	63	9.4	84
4	1.6	30	34	2.9	52	64	3.7	30
5	5.6	70	35	2.1	43	65	4.9	61
6	6.0	62	36	4.8	35	66	1.8	40
7	4.3	47	37	8.1	58	67	3.6	24
8	8.1	72	38	1.2	5	68	2.9	37
9	5.7	40	39	4.6	46	69	2.6	51
10	0.3	53	40	4.0	57	70	3.1	30
11	1.6	18	41	6.1	40	71	4.7	47
12	2.5	48	42	0.8	39	72	5.9	54
13	5.8	53	43	5.9	42	73	2.5	47
14	5.3	48	44	6.5	62	74	4.6	52
15	6.3	64	45	6.5	52	75	5.2	52
16	3.4	52	46	7.5	76	76	7.6	52
17	6.2	61	47	7.2	67	77	3.7	33
18	3.2	34	48	6.7	45	78	7.4	74
19	6.3	65	49	4.1	23	79	3.3	56
20	6.1	66	50	4.0	33	80	5.1	53
21	4.6	33	51	4.8	59	81	5.1	54
22	6.7	76	52	4.0	50	82	7.5	76
23	0.5	34	53	6.2	49	83	4.7	37
24	3.2	46	54	5.5	44	84	2.7	45
25	5.3	55	55	7.0	52	85	2.7	5
26	5.0	33	56	7.5	68			
27	4.8	46	57	6.2	48			
28	6.4	49	58	5.7	55			
29	0.2	17	59	7.3	77			
30	4.9	47	60	1.9	9			

## **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   39.00   50.00   48.69   59.00   84.00
```

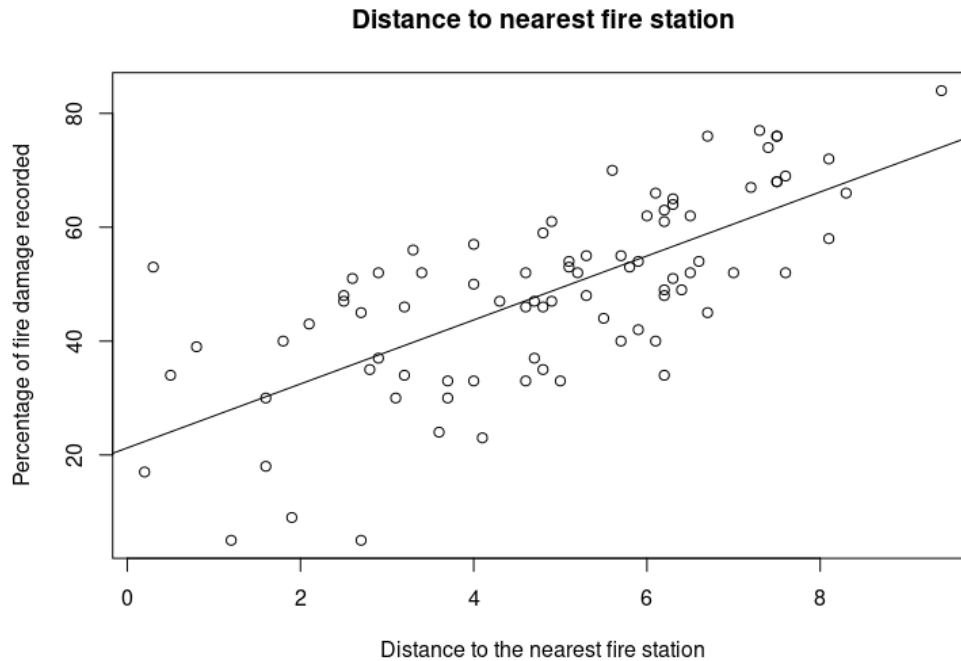
## **Exhibit 3**

```
> # Obtain residual values
> linear_model = lm(Percent~Distance)
> residual_values = resid(linear_model)
> # Summary statistics for the residual values
> summary(residual_values)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-24.523  -8.415   1.091   0.000   8.220   28.289
```

## **Exhibit 4**

```
> # Compute the sample correlation coefficient
> r <- cor(Distance, Percent)
> r
[1] 0.7108801
> # Carry out hypothesis testing
> n = 85
> test_statistic = r * sqrt(n-2)/sqrt(1-r^2)
> test_statistic
[1] 9.208453
> p_value = pt(test_statistic, n-2, lower.tail = FALSE)
[1] 1.249527e-14
```

**Figure 1**

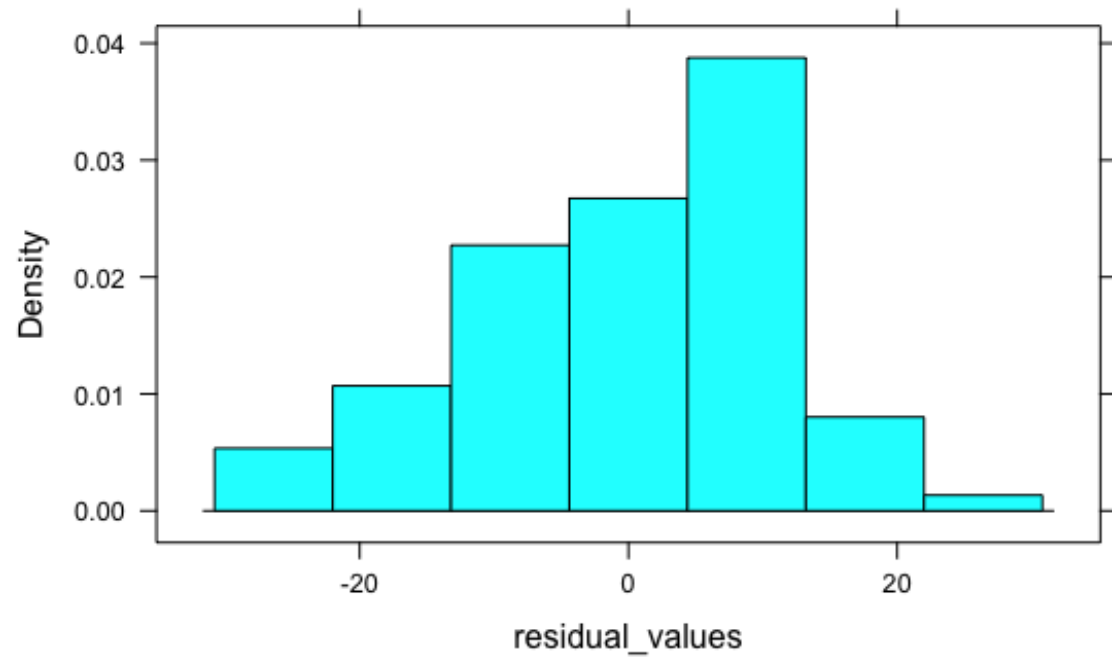


```
> plot(Distance, Percent, main = "Distance to nearest fire station", xlab =  
"Distance to the nearest fire station", ylab = "Percentage of fire damage  
recorded", abline(lm(Percent ~ Distance)))
```

```
> mod <- lm(Percent ~ Distance)  
> my_equation <- paste("y = ", coef(mod)[[1]], "+", coef(mod)[[2]], "* x")  
> my_equation  
[1] "y = 21.2381386939131 + 5.62080493982994 * x"
```

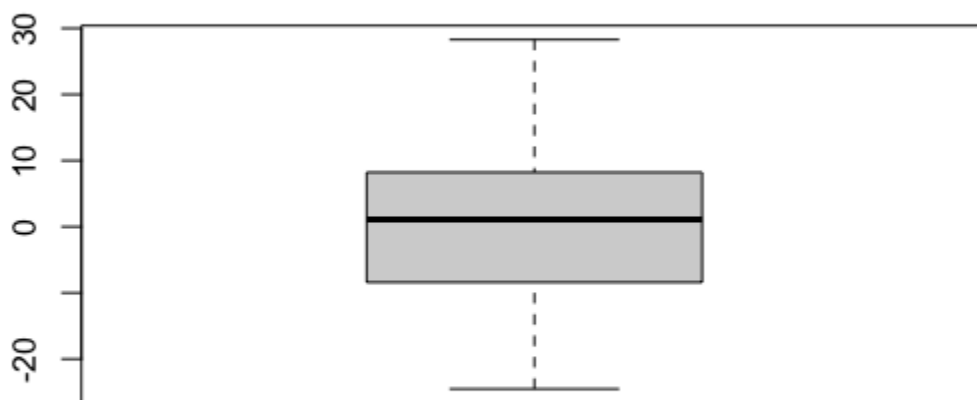
```
> cor(Distance, Percent)  
[1] 0.7108801
```

**Figure 2**

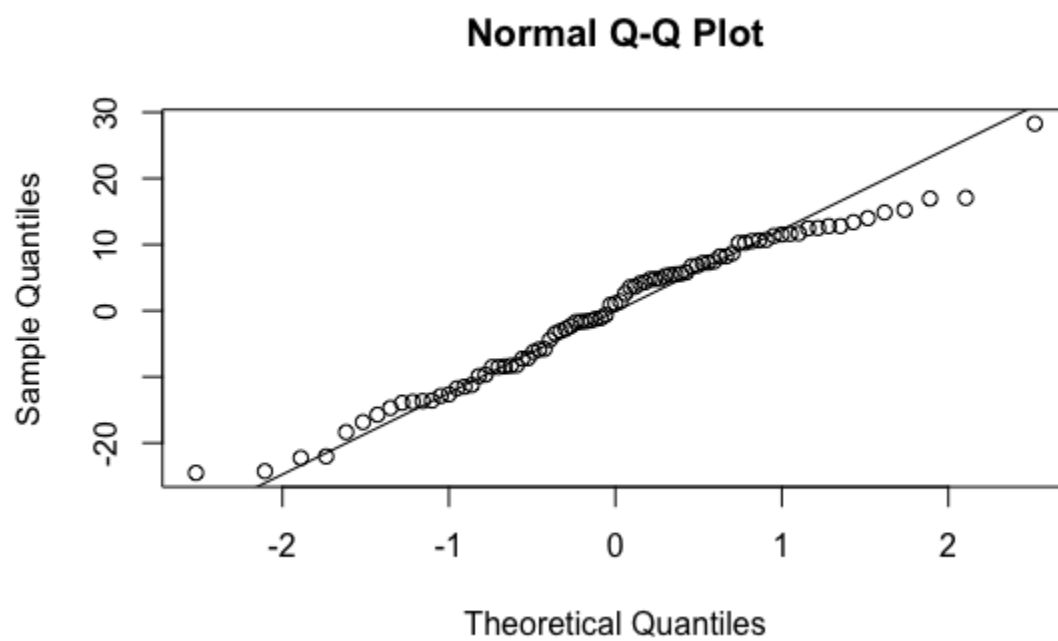


**Figure 3**





**Figure 4**



**Figure 5**

