Age and Insurance Premiums

How does age correlate with insurance pricing?

Mathematics

Word Count: 2668

Table of Contents

Introduction

Insurance is an important service to protect people from potential issues, disasters, or mistakes that occur later in life. There are many different types of insurances, varying from life insurance (a payout to your family in case of a death) to fire insurance (payout in the instance of a destructive fire). Usually, a person will pay a monthly fee, known as a premium, to retain coverage from an insurance provider, given that the provider will help pay for future repairs or damage as needed. The main type I will be focusing on in this essay is automobile insurance. Auto insurance covers the cost of any potential costs incurred while driving. When deciding on a premium, insurance companies will often hire actuaries to determine the risk of an individual applying for insurance. Many different factors can affect the cost of insurance such as location, annual mileage, credit record, and type of vehicle. The higher the risk of a person to crash, the higher the cost of their insurance.

The mathematical modeling of how insurance prices change based on certain factors is extremely important to determining the best price for insurance applicants. If a person is more likely to crash, it is beneficial to an insurance company to charge a high risk individual more money per month. If a person has a very good driving record, it is important for an insurance company to charge a low premium so the customer is likely to stay with the company.

Significance and Linear Regression

In the process of creating a model for age and insurance pricing, it is important to determine the significance of the discovered relationship between age and insurance pricing. To do this, the slope of the least squares regression line (LSRL) will be tested in a linear regression (LinReg) test. An LSRL can be used to make predictions of a general population based on samples.

1.1: Statistical Significance

A result can be defined as statistically significant when it is very unlikely to occur given a null hypothesis, H_0. When an alternative hypothesis (H_a) is tested against H_0, and the probability p is less than the given significance level (a), then H_0 can be rejected and H_a is true. Statistical significance is very important because it tells if a result can be generalized to a population. If the probability of a result is less than the significance level, it can be concluded that the sampling error is not responsible for the results and they can be generalized to the whole population and therefore more useful for predictions. If the probability is higher than the significance level, then there was not enough of a correlation to generalize from the sample.

1.2 Errors

There are two types of errors: type I error and type II error. A type I error occurs when the null hypothesis is true, but it is rejected. This denotes that a false positive has

occurred, that we assert a new claim that is not true when compared to H_0. To minimize type I error, an investigation would use a low significance level to limit the room for detecting a false positive. The other type of error, type II, occurs when H_0 is false but is not rejected. This would imply that a false negative has occurred, or that the results were statistically significant but the H_0 was not rejected. To minimize a type II error, a higher significance level should be used as it will allow for more statistically significant results to not be rejected.

Since type I and type II error require the opposite conditions to be minimized, the significance level needs to be chosen in the context of which error is worse should it occur. Choosing a higher significance level would imply that the type II error would be worse should it occur, while a low significance level implies that a type I error would be worse if it occured.

In the context of a company providing car insurance, a type I error would mean there is no relationship between age and annual price of insurance, but it is concluded that there is a relationship. A type II error would imply that there is a relationship between age and the price of insurance, but it is concluded that there is no relationship. Since a company is the most likely to be doing an analysis of age and annual price, the point of view of the company providing insurance will be used. For a theoretical company, a type I error is worse because it means that the company would be giving estimates that are largely inaccurate and based on the results that were due to random chance, hurting their reputation.To minimize this, a lower significance level of .05 will be used.

1.3 Linear Regression Test

A linear regression (LinReg) test determines the significance of the relationship between two variables. In a LinReg test, the slope or relationship found ($H_a : b_0 \neq 0$) is tested against no relationship ($H_0 : \ b_0 = 0$) where b_0 is the slope of a linear regression relationship. If the relationship is significant, this means that the equation found is useful for making predictions about ages that were not observed.

In order to do a LinReg test, certain conditions must be satisfied. The relationship between the independent variable x and the dependent variable y must be linear and random. The distribution of the dependent variable y must be normal or approximately normal and the standard deviation must be constant.

1.4 Equations

These three equations can be used to look for relationships within data:

$$\hat{y} = b_1 + b_0 x$$

$$b_0 = \frac{r S_y}{S_x}$$

$$b_1 = \bar{y} - b_0 \bar{x}$$

$\hat{y}$ is the predicted insurance annual cost

b_1 is the y-intercept

b_0 is the slope

S_y is the standard deviation of annual insurance prices

This can be calculated as:

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

n is the sample size

y_i is each recorded annual insurance price

S_x is the standard deviation of age

This can be calculated as:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

x_i is each recorded annual insurance price

r is the correlation coefficient

$\bar{x}$ is the average age

$\bar{y}$ is the average annual insurance cost

When finding the test statistic

$$t = \frac{b - \beta}{SE_b}$$

b is the slope of the sample

β is the slope of the null hypothesis

SE_b is the standard error of the sample slope

When finding the standard error of the sample slope

$$SE_b = \frac{s}{\sqrt{n}} \left(\frac{1}{S_x}\right)$$

When creating a confidence interval:

$$b \pm t^* \left(\frac{s}{\sqrt{n}}\right)$$

n is the sample size

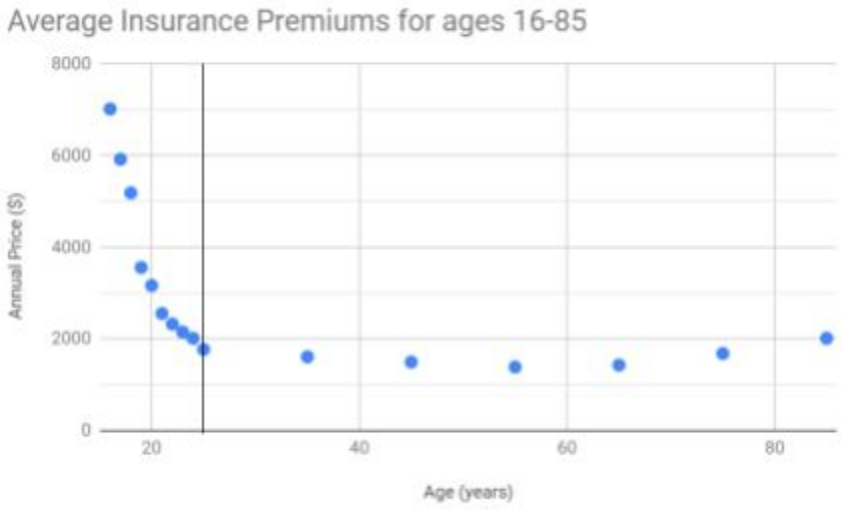
s is the sample standard deviation

t^* is the critical value

All above equations from *The Practice of Statistics*. Equation of standard error from "Mathematics of Simple Regression"

Methodology

I will create multiple LSRLs to determine the relationship between age and insurance pricing. Since the relationship between age and insurance is nonlinear, I will divide the ages into 2 categories: young (16-25) and old (35-85).



For both of these age groups, I will perform a LinReg test to determine the significance of the slope.I will interpret the meaning of the slopes for each age group and what use they could have.

Young Age Group (16-25)

In the 16-25 year old age group, there is a strong, negative, linear relationship between age and average insurance price. The LSRL between age and annual insurance prices can be found:

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{10}(y_i - 3567.620)^2}{10-1}}$$

$$S_y = \sqrt{\frac{30{,}503{,}371.68}{9}}$$

$$S_y = \sqrt{\frac{30{,}503{,}371.68}{9}}$$

$$S_y = 1840.995$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{10}(x_i - 20.5)^2}{9}}$$

$$S_x = \sqrt{\frac{82.5}{9}}$$

$$S_x = 3.028$$

$$b_0 = \frac{rS_y}{S_x}$$

$$b_0 = \frac{(-.938)(1840.995)}{(3.028)}$$

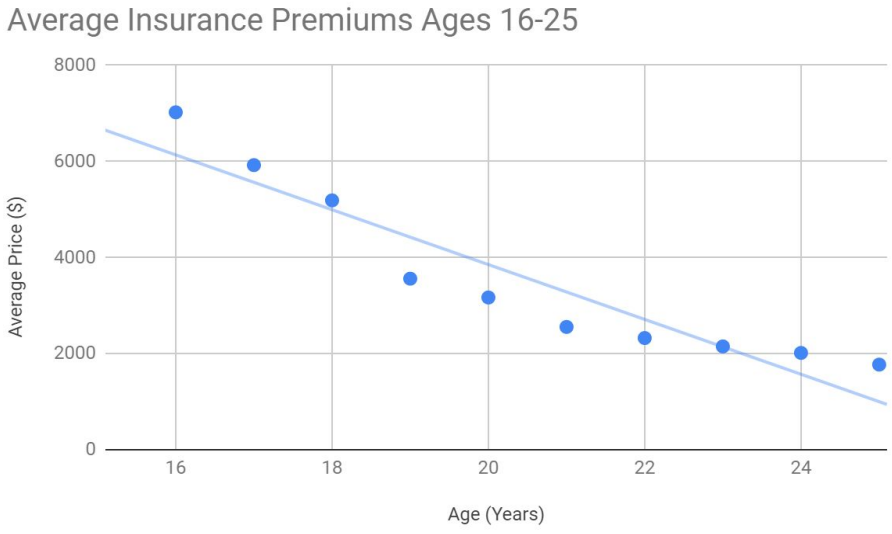
$$b_0 = -570.295$$

$$b_1 = \bar{y} - b_0\bar{x}$$

$$b_1 = 3567.620 + (570.295)(20.5)$$

$$b_1 = 15258.668$$

Which leads to an LSRL of $\hat{y} = 15268.668 - 570.295x$, as shown below with domain limit $16 < x \leq 25$.

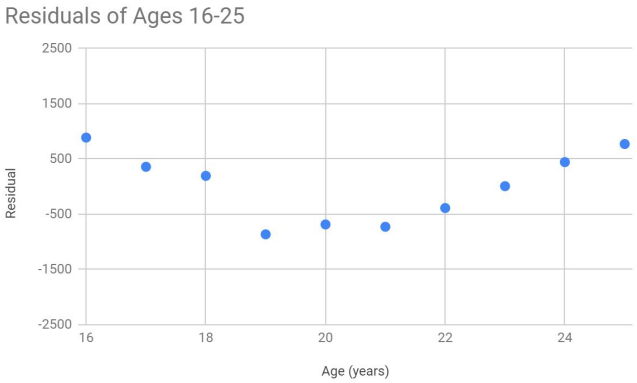


The correlation coefficient can be calculated as follows:

$$r = .938$$

$$r^2 = .8798$$

88.0% of variation in insurance pricing is explained by age. For each additional year of increased age, the predicted average cost of insurance decreases by about $570.30.



The residual plot shows a pattern as age increases, meaning the LSRL is not necessarily appropriate. This could be due to external factors affecting the price of insurance that are not related to age such as gender and type of vehicle. As there is not many as many factors to consider for younger individuals, age is acts as an important factor explaining car insurance pricing, reflected in the high r^2 value. Although the line is not necessarily accurate, a test will still be done, with the possible external factors kept in mind.

The null hypothesis is there is no relationship between age and insurance pricing ($H_0 = 0$). The alternate hypothesis is that there is a relationship between age and insurance pricing $(H_a \neq 0)$.

When calculating the probability of the slope being significant the standard error of the sample slope and the test statistic is found as:

$$SE_b = \frac{s}{\sqrt{n}}\left(\frac{1}{S_x}\right)$$

$$SE_b = \frac{674.591}{\sqrt{10}}\left(\frac{1}{3.028}\right)$$

$$SE_b = 70.451$$

$$t = \frac{b-\beta}{SE_b}$$

$$t = \frac{-570.295-0}{70.451}$$

$$t = -8.095$$

Using this test statistic as well as degrees of freedom value of 8 results in a p-value of 2.005×10^{-5}, or approximately 0, is found. Since the p-value is less than .05, we can reject H_0 and say that there is a relationship between age and the annual price of insurance.

Using the found LSRL, we can make predictions about the possible price of a person at a given age between 16-25.

$$\hat{y} = 15268.668 - 570.295x$$

$$\hat{y} = 15268.668 - 570.295(22)$$

$$\hat{y} = 15268.668 - 12,546.49$$

$$\hat{y} = 2722.178$$

In this example, we can predict that at age 22, a person would have an annual insurance premium of $2722.18. This is not the same as the average of the collected prices for 22 year olds ($2323.96) possibly because of external factors affecting the price, such as those discussed in the introduction.

Since the results are significant, creating a 95% confidence interval can help show the relationship between age and premium price. Using a critical value of 1.860 from a calculator, a 95% confidence interval can be created:

$$b \pm t^*(\frac{s}{\sqrt{n}})$$

$$-570.295 \pm 1.860(\frac{674.591}{\sqrt{10}})$$

$$-570.295 \pm 396.783$$

$$(-967.078, -173.512)$$

The 95% confidence interval means that we can be 95% certain that the true relationship between age and annual insurance premiums is between -967.078 and -173.512. Since -570.295 is within this interval, it is a possible relationship between age and annual price.

Another important pattern to note is the change seen in the average and standard deviation of the annual prices with increasing age. The average annual price

stays above $5000 until the age of 18, but is followed by a large fall to about $3500 at 19, and consistent decreases down to $1770.96 per year Standard deviation remains above $1400 until the age of 18, then drops blow $1000 at 19, then decreases down to $454.82. The largest drops in both of these values within this age group occur at the age where students are leaving high school and becoming adults, which could be a significant factor determining the annual price of insurance.

Old Age Group (35-85)

In the 35-85 age group there remains a strong, negative, linear relationship between age and annual insurance premiums. A LSRL is calculated below:

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{6}(y_i - 1602.330)^2}{5}}$$

$$S_y = \sqrt{\frac{263{,}924.498}{5}}$$

$$S_y = 229.750$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{6}(x_i - 60)^2}{5}}$$

$$S_x = \sqrt{\frac{1750}{5}}$$

$$S_x = 18.708$$

$$b_0 = \frac{rS_y}{S_x}$$

$$b_0 = \frac{(.611)(229.750)}{(18.708)}$$

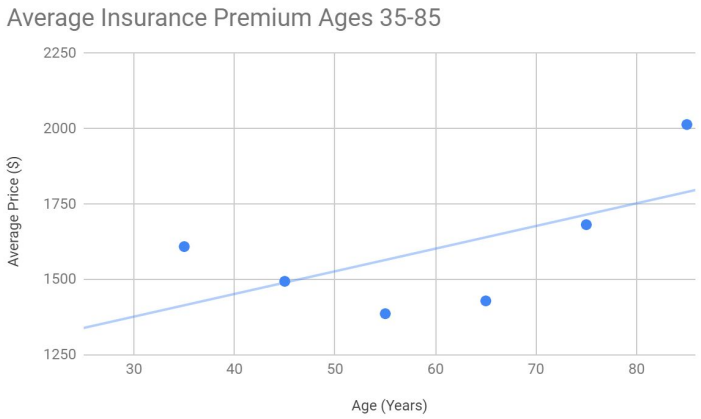
$$b_0 = 7.504$$

$$b_1 = \bar{y} - b_0\bar{x}$$

$$b_1 = 1602.330 - (7.504)(60)$$

$$b_1 = 1152.09$$

The resulting LSRL, $\hat{y} = 1152.09 + 7.504x$, is pictured below. For the purposes of remaining in the age group, the domain of the ages will be limited to $25 < x \leq 85$.

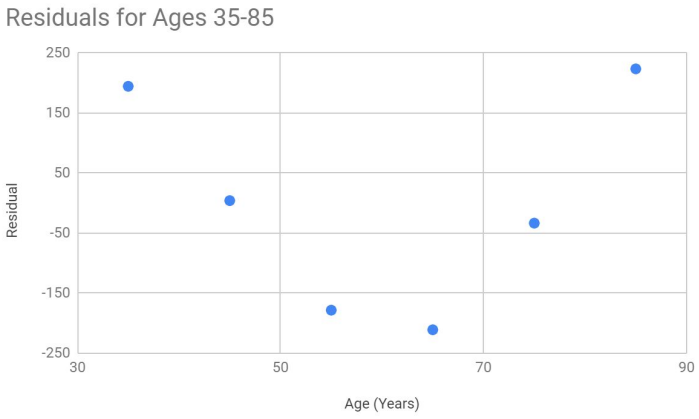


The correlation coefficient is as follows:

$$r = .661$$

$$r^2 = .373$$

37.3% of variation in insurance pricing is explained by age. For each additional year of increased age, the predicted average cost of insurance increases by about $7.504.



The residual plot shows a small pattern in the data, but since the data is limited a pattern is difficult to define. Just as in the young age group, this LSRL may not be appropriate because of external factors influencing the pricing of car insurance. This is more apparent in the older age group because there are more factors to consider when determining prices, such as the driver's record. A test and confidence interval will still be made with the LSRL in spite of these issues.

The null and alternate hypothesis will remain the same as in the young age group, $H_0 = 0$ and $H_a \neq 0.$ The standard error of the sample slope and the test statistic can be found in the same manner as previously used:

$$SE_b = \frac{s}{\sqrt{n}} \left(\frac{1}{S_x}\right)$$

$$SE_b = \frac{214.835}{\sqrt{6}} \left(\frac{1}{18.708}\right)$$

$$SE_b = 4.851$$

$$t = \frac{b - \beta}{SE_b}$$

$$t = \frac{7.504 - 0}{4.851}$$

$$t = 1.547$$

Using this test statistic and degrees of freedom value of 4, a p-value of .197 is obtained. Since the p-value is greater than the significance level of .05, we can not reject H_0 and do not say that there is a relationship between age and the annual price of insurance.

One reason that the results from the older age group are not significant is because of the greater variety of factors influencing their price. A given individual's driving record, financial situation, or credit score might all have an unforeseen impact on the price of insurance. Because of the high probability that the results are due to random chance, it is not appropriate to make predictions with this model.

Conclusion

Overall, the equation to predict the price of insurance based on age was more accurate and useful for individuals between ages 16 and 25 while age did not act as a good predictor for price on individuals in the older age bracket. This is largely due to the increase in confounding variables affecting the data of older individuals, such as their driving record and credit score.

Although the older age model is not the best for making a price prediction, The model for younger drivers could be used by a car insurance company in order to give a monthly premium price to individuals between ages 16 and 25.

As type I error is worse for a company providing car insurance, I used a smaller significance level during my analysis. This was to minimize the likelihood that a car company would use these results to make predictions when the results might have been based on random chance. Although a type II error fails to provide a potentially usable model for a car company, the potential of using an incorrect model poses a much greater concern to a company than having to come up with a new model.

It is interesting to note that the results with the higher correlation coefficient yielded a lower p-value. Further investigation into the relationship between correlation coefficient and p-value may prove to be useful information.

Works Cited

"Average Car Insurance Rates by Age." *CarInsurance.com*, QuinStreet Insurance Agency, 27 Nov. 2018, www.carinsurance.com/average-rates-by-age.aspx. Accessed 9 Dec. 2018.

Nau, Robert. "Mathematics of Simple Regression." *Statistical Forecasting: Notes on Regression and Time Series Analysis*, Duke University, people.duke.edu/~rnau/mathreg.htm. Accessed 5 Nov. 2019.

Starnes, Daren S., et al. *The Practice of Statistics*. 5th ed., 2015.