**Introduction:**

Health Insurance being a volatile field due to all the internal and external changes occurring continuously, requires a hefty load of analysis to understand the impacts of several factors individually and correlatively with one another to understand and draw conclusions with the end goal of building a product that provides sufficient coverage with a reasonable premium.

The Problem Statement is a watered-down scenario of the real-life application of Data Analysis, Manipulation & Interpretation in the field of Insurance, where a health insurance company is currently analysing the Impact of Gender and Time on the response variable of Claim Frequency and Claim Count.

A number of tests are conducted to fully understand the relationship between different variables, as well as their effect on the resulting estimated values.

**Confidence Interval:**

A Confidence Interval is a technique of estimating the value of an unknown parameter in the form of an interval with stated probability, where the width of the interval is the measure of the accuracy of the estimator.

To Check if the Average Claim Count for Males is higher than Females, The Confidence Interval is presented in the form of 2 Sample t-Test (as the population variances are unknown but assuming both the variances are equal), where is supposed to be the ideal scenario.

x̅m = 159.7333 x̅f = 153.133

sm = 44.6342 sf = 41.4881

nm = 15 nf = 15

Pivotal Quantity:

The 100(1-α)% Confidence Interval for is given by,

Where,

= 1856.737126

Therefore, the 95% Confidence Interval for is,

=

=

The Confidence Interval Includes 0 in its range, which suggests that the results are not statistically significant,

And the lower bound < 0, this suggests that there is quite the possibility that Average Claim Cost for males might be lower than Females.

Therefore, there is not sufficient evidence to conclude that the average claim count for males is higher than females.

**Hypothesis Testing:**

Hypothesis Testing tests an assumption (in this case, the assumption being ). This Assumption is often called a Null Hypothesis or , and is tested against an Alternative Hypothesis or , to check the plausibility of the Base Assumption.

Similar to the Confidence Interval, a 2-sample t-Test is conducted to test the Null Hypothesis.

A test to check if the Claim Costs for Male is equal to that of Females.

vs

Test statistic:

This is a two-sided t-test and our test statistic is less than 2.048 (the lower 2.5% point of the distribution) so we have insufficient evidence to reject at the 5% level.

Therefore, we conclude that the claim costs for male is equal to that of females.

P – value:

this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is:

By interpolating, we get p-value as

P – value

Since this is greater than 0.05, we have insufficient evidence to reject at the 5% level. And thus, having sufficient evidence to reject at 5% level.

Hence, we conclude that the claim costs for male is equal to that of females.

**Scatter plot:**

|  |  |
| --- | --- |
| Year | Claim Count(M) |
| 2010 | 100 |
| 2011 | 97 |
| 2012 | 114 |
| 2013 | 122 |
| 2014 | 130 |
| 2015 | 139 |
| 2016 | 148 |
| 2017 | 157 |
| 2018 | 153 |
| 2019 | 177 |
| 2020 | 188 |
| 2021 | 199 |
| 2022 | 211 |
| 2023 | 224 |
| 2024 | 237 |

Here we can see that there appears to be a strong positive linear relationship. The plotted data points lie roughly in a straight line.

You can see from the graph that there appears to be a linear relationship between the year and claim counts of males (i.e as the year increases the rate of claim counts of males are also increasing).So we will be able to use our linear regression work on these data values.

|  |  |
| --- | --- |
| Year | Claim Count(F) |
| 2010 | 97 |
| 2011 | 102 |
| 2012 | 110 |
| 2013 | 115 |
| 2014 | 130 |
| 2015 | 132 |
| 2016 | 138 |
| 2017 | 140 |
| 2018 | 156 |
| 2019 | 169 |
| 2020 | 176 |
| 2021 | 191 |
| 2022 | 197 |
| 2023 | 220 |
| 2024 | 224 |

Here we can see that there also appears to be a strong positive linear relationship just like the male claim counts and similarly, the plotted data points roughly form a straight line.

You can see from the graph that there appears to be a linear relationship between the year and claim counts of females (i.e as the year increases the rate of claim counts of females are also increasing).So we will also be able to use our linear regression work on these data values.

**Linear Regression:**

The Linear Regression Model is the simplest and one of the foundational models in the field of estimation. It estimates the linear relationship between variables and response. It is usually formatted as , where is the intercept or bias, and is weight of the variable

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **SUMMARY OUTPUT** |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | *Regression Statistics* |  |  | |  | | --- | |  | |  |  |  |  |  |  |  |  |  |  |
|  | Multiple R | 0.005169385 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | R Square | 0.000026723 |  | |  |  |  |  |  |  |  |  |  |  |
|  | Adjusted R Square | 0.035686609 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Standard Error | 169.9319079 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Observations | 30 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **ANOVA** |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |  |  |  |  |  |
|  | Regression | 1 | 21.60714286 | 21.60714286 | 0.000748251 | 0.978371276 |  |  |  |  |  |  |  |  |
|  | Residual | 28 | 808551.8929 | 28876.85332 |  |  |  |  |  |  |  |  |  |  |
|  | Total | 29 | 808573.5 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |  |  |  |  |  |
|  | Intercept | 1542.303571 | 14483.9766 | 0.106483434 | 0.915958066 | -28126.77753 | 31211.38468 | -28126.8 | 31211.38 |  |  |  |  |  |
|  | X Variable 1 | 0.196428571 | 7.180933748 | 0.027354182 | 0.978371276 | -14.5130474 | 14.90590455 | -14.513 | 14.9059 |  |  |  |  |  |

**y** = 1,542.30 + 0.1964**x**   
R² = 0.000026723

* The model suggests that with passing years the claim cost increases by a unit of 0.1964. Additionally, there is a fixed cost of 1542.30 incurred every year.
* Coming to the statistical significance of the parameters opted for this model. P-value or probability value is a number which helps interpret to what extent the dependent and independent variables are associated

A low P-value (<0.05 for 95% confidence interval) indicates high association whereas a high P-value indicates there is less meaningful association between the variables

* Here a P-value =0.978371276 strongly suggests that year is not a strong regressor to predict the claim costs
* Another parameter for judging the credibility of a model is the Coefficient of Determination. This tells us how much variability is explained by the model. Thus, a High R-squared value is preferred

However, for this model the R-squared value=0.000026723 which us that the model does not fit the data well and hence does not predict the regressed variable correctly.

**Use of Generalised Linear Model:**

In the summary statistics for the simple linear regression model, we observed that the coefficient of determination (R²) for the fit between time and claim cost is close to 0, which suggests that the model does not explain a significant variability in the response variable. This outcome may indicate either a poor model fit or violations of the underlying assumptions, particularly the assumption of normality of the residuals.

To address these potential issues, a Generalized Linear Model (GLM) presents a more flexible approach compared to traditional linear regression. GLMs extend the basic linear model framework by allowing the response variable to follow distributions other than the normal distribution, thereby accommodating a broader range of data types and relationships. This makes GLMs particularly useful in situations where the assumptions of normality and constant variance are violated.

Additionally, gender as an attribute, typically a categorical variable, can also be effectively incorporated into a GLM structure. GLMs allow for the inclusion of categorical predictors with appropriate link functions, making it possible to model relationships between such variables and the outcome are more aligned than simple linear regression.

Hence variables such as claim count and claim cost, which are typically suited for a Poisson or Gamma distribution, a GLM would likely provide a more appropriate model. The Poisson distribution is often used for discrete data hence Claim Count as the response is well regressed on gender and time, while the Gamma distribution is suitable for modelling continuous, positive-valued with skewed distributions thus Claim Cost is well regressed using this distribution.

On conversion of simple linear regression models to GLM using proper link functions we got an adjusted R^2 value close to 99%. One limitation of using excel is the lack of functionalities for GLM hence parameters like Akaike’s Information Criteria (AIC), Bayesian Information Criteria (BIC) are not available for the purpose of comparisons.

In conclusion, the use of a Generalized Linear Model would offer a more robust and flexible modelling approach for our dataset, accommodating the potential non-normality of the response variables and providing a better fit for the underlying data.