Brad Bohacs

Regression Student Projected

Submitted 01/20/2016

**Introduction:**

The cost of healthcare continues to rise in the US over double the rate of CPI. Companies that sponsor health benefit plans are looking for every way possible to slow the rate of healthcare increase. Companies ask all the time what influences an employee's healthcare cost, with the hopes of isolating a few key factors and ultimately put programs in place to mitigate the healthcare cost. This exercise looks at employees healthcare cost on an individual basis and analyzes the demographic information available to understand what demographic factors influence healthcare cost.

Employee level healthcare cost, which included point of service claims cost and biweekly contributions paid to the employer to be enrolled in the health plan, were used to perform a regression on healthcare cost with the following explanatory variables: employees salary, gender, age, plan choice (PPO or CDHP), enrollment in a dental plan, and tobacco user status.

Regression statistics were used to measure the validity and accuracy of the regression explanatory demographic factors on the cost of healthcare, the regressor. This includes looking at the regression equation, total sum of squares, regression sum of square, residual sum of squares, R^2, adjusted R^2, the T test at the null hypothesis, the p value for the T Test, the F test at the null hypothesis, and the significance of the F test

**Data:**

1,519 employee records were used that had calendar year 2014 data. The only data available includes all employees enrolled in the healthcare plan. Employees not enrolled in the healthcare plan were removed due to the 0's they would produce; we only care about the cost of healthcare and don't want the employees with no cost influencing the analysis. Employees that were not enrolled in the healthcare plan for 12 months were also removed from the analysis to remove seasonality.

All the data is from the employer except the healthcare claims cost which was from the medical/Rx vendor. Data was adjusted to not be identifiable.

To assist with scaling issues, salary and healthcare cost was adjusted to represent dollars in thousands. This made the coefficient and regression model more comparable to the scale of age (the older person in the data set is 65). Gender, plan choice, enrollment in the dental plan, and tobacco use status were identified as dummy variables since they are qualitative (can only be yes or no) instead of quantitative. Here is a list of how the dummy variables were assigned:

- Gender: Male was assigned a 1, female was assigned 0
- Plan design: PPO was assigned 1, CDHP was assigned 0
- Enrolled in Dental: yes was assigned 1, waiving dental coverage was assigned 0
- Tobacco user: A tobacco user, which was self-reported, was assigned a 1 while a user that responded no was assigned a 0.

**Regressions Equation:**

HealthCare Cost in Thousands (Fitted Value) = $5.93 + (0.007 * Salary) + (.093 * Male ) + (-0.015 * Age) + (1.38 * PPO) + (2.089 * Enrolled in Dental) + (-0.468 * Tobacco User)

A brief explanation of what each measure means:

- Intercept: $5.93 represents the baseline cost. If all other explanatory values are 0, the fitted value of healthcare cost in thousands would be $5.93
- Salary Coefficient: For every $1,000 change in salary, with all other things being equal, the fitted value of healthcare cost in $1,000s is expected to change 0.007. This is a direct relationship, as salary increases so does healthcare.
- Sex Coefficient: If the employee is male, with all other things being equal, the fitted value healthcare cost in $1,000s is expected to change 0.093.
- Age Coefficient: For every change in 1 year of age, with all other things being equal, the fitted value of healthcare cost in $1,000s is expected to change -0.015. This is an inverse relationship, as age increases the cost of healthcare decreases.
- Plan Coefficient: If the employee is enrolled in the PP0, with all other things being equal, the fitted value healthcare cost in $1,000s is expected to change 1.38
    - This is somewhat expected due to the greater cost in employee biweekly contributions of being enrolled in the PPO
- Dental Plan Coefficient: If the employee is enrolled in a dental plan, with all other things being equal, the fitted value of healthcare cost in $1,000s is expected to change 2.089
- Tobacco User Coefficient: If the employee self-identified as a tobacco user, with all other things being equal, the fitted value of healthcare cost in $1,000s is expected to change 0.093

**Statistical Techniques:**

Total Sum of Square

The total sum of squares measures the total variance of a distribution. It does not change with the regression model; a data set has a fixed total sum of squares. It is calculated as the sum of the square difference between the actual health care cost and its mean. The total sum of squares for this data set is 18,420. In the regression model, the total sum of squares can be broken down into two components, the regression sum of squares and the residual sum of squares

Regression Sum of Squares

The regression sum of squares measures the amount of the total variance accounted for by the regression model. It is calculated as the sum of the square difference between the fitted healthcare cost (using the regression model) and the mean of healthcare cost. For the regression model, this is 1,138.

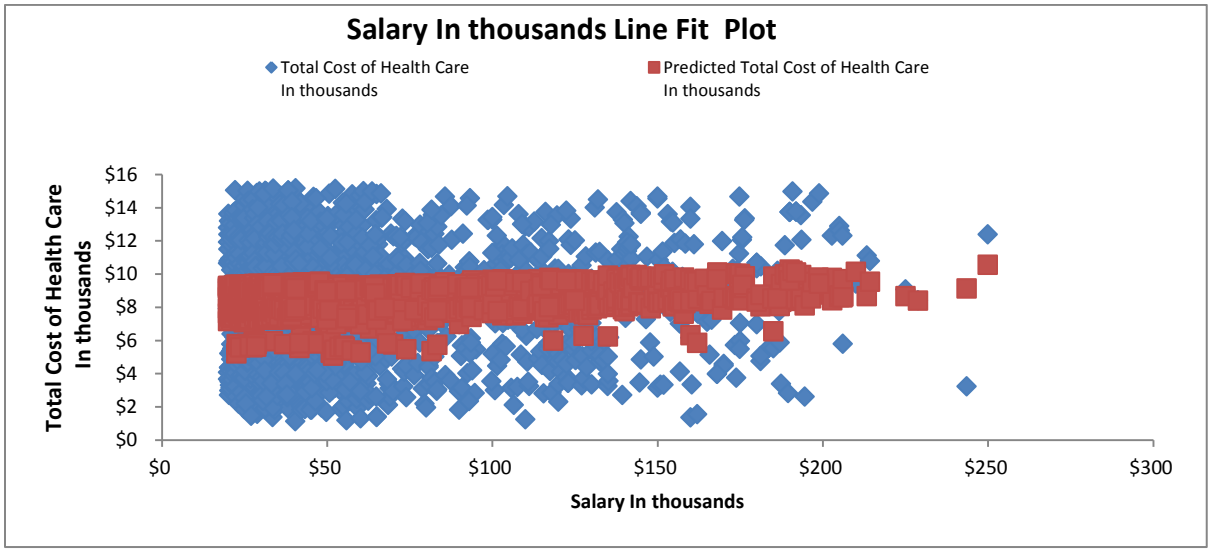
Residual Sum of Squares

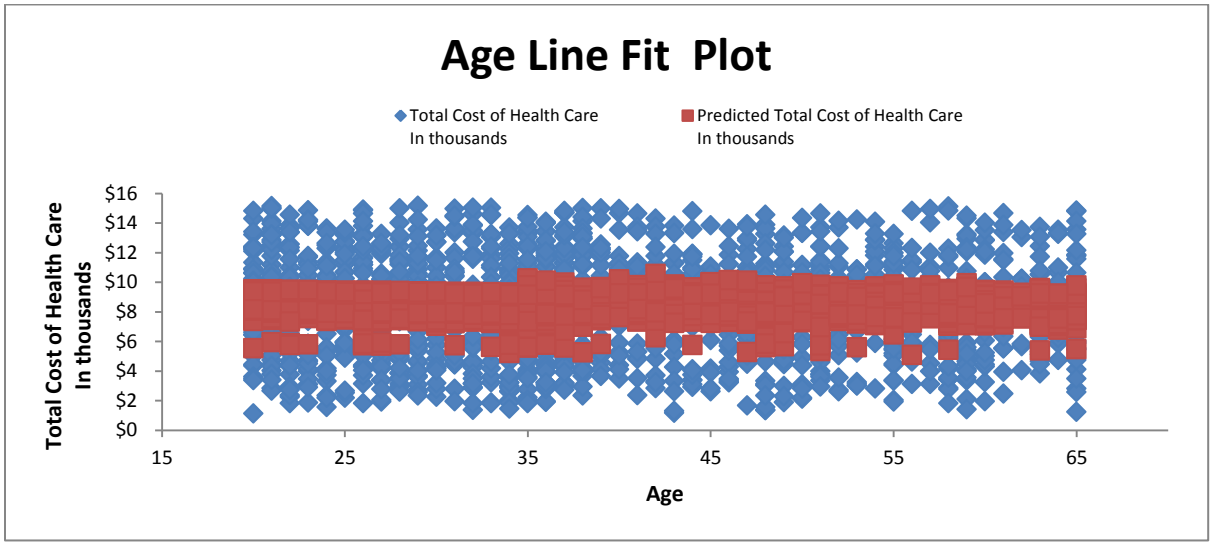
The remaining 17,282 is the residual sum of squares. It measures the amount of variance not accounted for in the regression sum of squares. It is calculated as sum of square difference between the fitted healthcare cost mentioned above and the actual cost of healthcare

R^2

R^2 is designed to measure the % of variance measured by the regression model, which is calculated by the regression sum of squares over the total sum of squares. For our model, R^2 is 6.178%, meaning our model only explained ~6.2% of the data's variance. This is not high at all, which can be explained by the two charts below. The blue dots are the actual cost of healthcare and red dots are the predicted amount.

As we can see, the blue dots are all over the place and well above and below the lines when we plot healthcare cost on both salary and age. Since the blue dots are far from the line, and there are plots far above and below the line for similar explanatory values, we get a high total sum of squares. Our model does not have an explanatory variable that can explain that variance at variance levels of salary and age.



Adjusted R^2

Adjusted R^2 is R^2 adjusted for the number of variables in the model. Naturally, R^2 will increase with the number of regressors added to the model. Adjusted R^2 adjusts for the number of variables by using a percentage of your sample size less number of predictors to the total sample size. For this model, the adjusted R^2 is 5.8%. This is useful when comparing multiple models, which will be discussed later.

Standard Error:

An additional measure of our regression equation is the standard error. This measures on average, how wrong our regression model is. It is calculated as the residual sum of squares (which as we mentioned is the sum of the squares of the actual Y minus the fitted Y) divided by the # of entries adjusted for the number of coefficients (in this case it is 1512). The standard error for our regression equation is 3.38. This means the fitted line is on average 3.38 thousands of healthcare cost away from the actual.

T Test and P test

The T-test was evaluated at the null hypothesis (that each coefficient is 0 and has no significant bearing on the model). This simply evaluates whether each variable is statistically different than 0, by evaluating how many standard deviations each coefficient is from 0. Surprisingly, gender and age are not statistically different than 0 and we do not reject the null hypothesis. For all other measures we reject the null hypothesis.

The T test was measured at 5% significance null hypothesis rejecting. Even if we reduced 2.5% significance to 2.5%, we would reject the same variables.

Below are the results of the analysis of the T test. 2 was used due to it being close to the 5% significance level with 1518 degrees of freedom.

| | ***t Stat*** | ***Absolute Value of T*** | ***Statistically Different than 0***<br><br>***Rule of Thumb is 2*** | ***Reject Null Hypothesis*** |
|---|---|---|---|---|
| **Salary** | 2.808874433 | 2.808874433 | Yes | Yes |
| **Gender**<br>**1 = Male**<br>**0 = Female** | 0.538179752 | 0.538179752 | No | No |
| **Age** | -1.781199851 | 1.781199851 | No | No |
| **Plan Design**<br>**1 = PPO**<br>**0 = CDHP** | 7.850977303 | 7.850977303 | Yes | Yes |

| | | | | |
|---|---|---|---|---|
| **Enrolled in Dental** | 4.409821542 | 4.409821542 | Yes | Yes |
| **Tobacco User** | -2.154399521 | 2.154399521 | Yes | Yes |

Next we measure the P value, which measures the distribution using 1,518 degrees of freedom. The p value was measured at 5% significance. If the p value is greater than 5%, we do not reject the null hypothesis

| | ***t Stat*** | ***P-value*** | ***Statistically Significant at 5%*** | ***Reject Null Hypothesis*** |
|---|---|---|---|---|
| **Salary** | 2.808874433 | 0.005035454 | No | Yes |
| **Gender**<br>**1 = Male**<br>**0 = Female** | 0.538179752 | 0.590532168 | Yes | No |
| **Age** | -1.781199851 | 0.075080503 | Yes | No |
| **Plan Design**<br>**1 = PPO**<br>**0 = CDHP** | 7.850977303 | 7.76212E-15 | No | Yes |
| **Enrolled in Dental** | 4.409821542 | 1.10765E-05 | No | Yes |
| **Tobacco User** | -2.154399521 | 0.031366408 | No | Yes |

F TEST

The F test Is similar to the T test but measures whether the overall regression model against a null hypothesis (a straight line intercept model). For this model, the F test is 16.59. When measure the significance of F, measured at 5% significance level, it is very small and therefore we reject the null hypothesis that the model is no better than a straight line intercept model.

**Conclusion**

In conclusion the regression model is not a good fit on predicting the cost of healthcare. I believe there are data points, not available to me, that do a better job at predicting the cost of healthcare. I do believe removing some of the demographic explanatory variables would enhance the model.

A regression was run using the available data without gender and I believe it is more accurate. This is due to the following stats:

| | All Data Available | Remove Gender (New) | Comments |
|---|---|---|---|
| **Total Sum of Squares** | 18,421 | 18,421 | Sum of squares is the same due to the same data set |
| **Reg. Sum of Squares** | 1,138 | 1,135 | Regression sum of squares under new model is slightly smaller when gender is removed due to less coefficients being used in the new model |
| **Residual Sum of Square** | 17,283 | 17,286 | Residual of new model is larger due to reg. sum of squares being smaller |
| **R^2** | 6.178% | 6.160% | R^2 is smaller due to reg. sum of squares being smaller in the new model |
| **Adj. R^2** | 5.806% | 5.850% | Adjusted R^2 is slightly larger in the new model. Adjusted makes the adjustment for the new number of coefficients. We want a model that is a simple as possible (fewest amount of coefficients), explaining the most of the variance. |
| **Standard Error** | 3.381 | 3.380 | Standard error is slightly smaller under the new model, meaning our regression line is slightly closer to the actual under the new model |
| **F Test** | 16.59365418 | 19.8637837 | F test under the new model is slightly larger which is good. We are further away from rejecting the null hypothesis |

The adjusted R^2, standard error, and F test favoring the new regression that removes gender, which leads me to my recommendation of using the new model

**Final regression model removing gender:**

HealthCare Cost (Fitted Value) = $5.99 + (0.0066 * Salary) + (-0.0147 * Age) + (1.38 * PPO) + (2.087 * Enrolled in Dental) + (-0.469 * Tobacco User)