Statistical Learning Models Project 3

Pedro Nunes

January 2024

1 Introduction

In this project, we'll explore simple and multiple logistic regressions using the Project 2 dataset.

2 Part 1 - Simple logistic regression model

2.1 Transformation of the Y-variable into a Categorical Variable

On Project 2, the Y-variable was the amount of money a company should charge for insurance, based on age, sex, children, smoker or not, and the region.

For the logistic regression, a qualitative variable is needed. Thus, I decided to calculate the mean of charges in the dataset and use it as a threshold. Every charge above the median would be a 1 (Above Average Charge), and every charge below would be a 0 (Below Average Charge).

Our data looked like this before changing the dependent variable (charges) to a qualitative one.

```
> head(data)
# A tibble: 6 × 7
    age   sex   bmi children smoker region charges
  <dbl> <dbl> <dbl>    <dbl>  <dbl>  <dbl>   <dbl>
1    19     1  27.9        0      1      0  16885.
2    18     0  33.8        1      0      1   1726.
3    28     0  33          3      0      1   4449.
4    33     0  22.7        0      0      2  21984.
5    32     0  28.9        0      0      2   3867.
6    31     1  25.7        0      0      1   3757.
```

Figure 1: Original data

And after transforming it into a qualitative variable and removing the charges from the dataset.

```
> # Display the modified data
> head(data)
# A tibble: 6 × 7
    age   sex   bmi children smoker region insurance_class
  <dbl> <dbl> <dbl>    <dbl>  <dbl>  <dbl>           <dbl>
1    19     1  27.9        0      1      0               1
2    18     0  33.8        1      0      1               0
3    28     0  33          3      0      1               0
4    33     0  22.7        0      0      2               1
5    32     0  28.9        0      0      2               0
6    31     1  25.7        0      0      1               0
```

Figure 2: Data with our new qualitative dependent variable

2.2 Getting the most correlated independent variable

```
> # Assuming 'data' is your dataframe
> correlation_matrix <- cor(data)
> correlation_matrix
                        age          sex          bmi    children       smoker       region insurance_class
age              1.000000000  0.020855872  0.109271882  0.04246900 -0.025018752 -0.002127313      0.15687113
sex              0.020855872  1.000000000 -0.046371151 -0.01716298 -0.076184817  0.004588385     -0.02835458
bmi              0.109271882 -0.046371151  1.000000000  0.01275890  0.003750426 -0.157565849      0.04170452
children         0.042468999 -0.017162978  0.012758901  1.00000000  0.007673120 -0.016569446      0.02423584
smoker          -0.025018752 -0.076184817  0.003750426  0.00767312  1.000000000  0.002180682      0.74625080
region          -0.002127313  0.004588385 -0.157565849 -0.01656945  0.002180682  1.000000000      0.03731662
insurance_class  0.156871130 -0.028354580  0.041704518  0.02423584  0.746250799  0.037316623      1.00000000
```

Figure 3: Correlation Matrix

For the simple logistic regression we want to use the most correlated variable. In this case, it is the smoker variable, with a correlation of 0.746.

2.3 Run logistic regression model with the most correlated variable

To do a logistic relation, we run the following code. Insurance_class is our dependant variable, while smoker is our independent class.

```
logistic <- glm(insurance_class ~ smoker, data=data, family="binomial")
```

Figure 4: Logistic regression

After, we can get details of our logistic regression, by running the summary command

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3506  -0.5453  -0.5453   0.0855   1.9897

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.83067    0.08884 -20.606  < 2e-16 ***
smoker       7.44015    1.00575   7.398 1.39e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.97  on 1337  degrees of freedom
Residual deviance:  867.84  on 1336  degrees of freedom
AIC: 871.84
```

Figure 5: Summary of logistic regression

From this, we can interpret the following model
charges = -1.83067 + 7.44015 * the person smokes
We can also conclude that

- **log(odds)** that a person who doesn't smoke pays an above median charge is -1.83067
- **log(odds ratio)** is the odds that a smoker will have to pay an above median charge over the odds that a non-smoker has to. Its value is 7.44015

The p-values are bellow 0.05, so our log(odds) and log(odds ratio) are statistically significant.

2.4 Get the estimate of the parameter in front of X

The estimate parameter of X is 7.44015. This suggests that a smoker is associated with an increase in the log odds of the event.

2.5 Calculate the predictor of p(Xi)

The P(X) formula can be given by

$$p(X) = \frac{1}{1 + e^{-\text{logit}(p(X))}}$$

Taking the formula into consideration, and replacing the logit function we used above we have,

$$p(X) = \frac{1}{1 + e^{-(-1.83067 + 7.44015 \times \text{smoker})}}$$

So, for a smoker, the probability of paying an above average charge is
P(1) = 0.9963504%
Whereas for a non smoker it is
P(0) = 0.1381579%

2.6 Explain the meaning of the predictor of p(Xi)

The value of P(X) ultimately gives you the probability of a smoker vs a non smoker having to pay an above average charge (Y=1).

3 Part 2 - Multiple logistic regression model

3.1 Run the Multiple Logistic Regression Model with project 2 chosen variables

During the project 2 the independent variables chosen were - age, sex, children, smoker and region.

Running the command glm(insurance_class . - bmi, data=data, family="binomial") we get

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.13357  -0.55242  -0.29890   0.04901   2.90706

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.580650   0.466178 -11.971  < 2e-16 ***
age          0.072099   0.008131   8.868  < 2e-16 ***
sex          0.265272   0.188232   1.409   0.1588
children     0.123953   0.075391   1.644   0.1001
smoker       8.394893   1.024868   8.191 2.59e-16 ***
region       0.180770   0.083913   2.154   0.0312 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.97  on 1337  degrees of freedom
Residual deviance:  759.56  on 1332  degrees of freedom
AIC: 771.56
```

Figure 6: Multiple logistic regression

By evaluation of the p-values, we can conclude that age, sex, smoker and region are all good estimators.

3.2 Get the estimates of parameters in front of each independent variable

By looking at the summary of our multiple logistic regression, we have that

- **age** 0.072099
- **sex** 0.265272
- **children** 0.123953
- **smoker** 8.394893
- **region** 0.180770

3.3 Compare the results with Part 1. Do you see any differences ?

We can see that the coefficient of smoker is bigger in the second part when compared to the first part.

3.4 Interpret the values of the parameters

- **age** older individuals are associated with a very low likelihood of paying an above average charge.
- **sex** females have a bigger likelihood of paying an above average charge.
- **children** is associated with a slightly bigger likelihood of paying an above average charge.
- **smoker** smokers have a very big likelihood of paying an above average charge
- **region** is associated with a slightly bigger likelihood of paying an above average charge

3.5 Calculate the predictor of p(X) in this model

As in part 1, p(x) is given by

$$p(X) = \frac{1}{1 + e^{-\text{logit}(p(X))}}$$

Replacing with the logit formula we got, we have that

$$p(X) = \frac{1}{1 + e^{-(\text{-5.580650}+0.072099\times\text{age}+0.265272\times\text{sex}+0.123953\times\text{children}+8.394893\times\text{smoker}+0.180770\times\text{region})}}$$

Using a example of a person, X, with

- **age** 30
- **sex** 1
- **children** 2
- **smoker** 1
- **region** 2

We have that

$$P(X) = 99.71$$

3.6 Explain the meaning of the number you have obtained

For the value of P(X) = 99.71% we have that a person with said features has a 99.71% chance of having an above average charge.

4 Conclusion

Overall our model is performing very well. The second model has better AIC (771.56), whereas the first model has an AIC of 871.84. Additionally it has a lower residual deviance (759.56) when compared to the simple logistic regression (867.84).

Additionally, we can calculate the R-squared value of our multiple logistic regression, following the formula

$$R^2 = \frac{\text{null deviance} - \text{proposed model deviance}}{\text{null deviance}}$$

- **null deviance** can be extracted from our logistic by running logistic$null.deviance/-2.
- **proposed model deviance** can be extracted from our logistic by running logistic$deviance/-2

and we get that

$$R^2 = 0.5437984$$

And we can aditionally use the same log-likelihoods to calculate a p-value for that R-squared using a chi-squared distribution, by using the formula

$$\text{Likelihood Ratio Test Statistic} = -2 \times (\text{Log-Likelihood of Null Model} - \text{Log-Likelihood of Proposed Model})$$

The p-value is then calculated using the chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the null and proposed models:

$$\text{p-value} = P(\chi^2 > \text{Likelihood Ratio Test Statistic})$$

In R we just need to run

1 - pchisq(2*(ll.proposed - ll.null), df=(length(logistic$coefficients) - 1))

Which gives us a p-value of 0, making our R-squared value statistically significant.

In the end, we get the following graphic

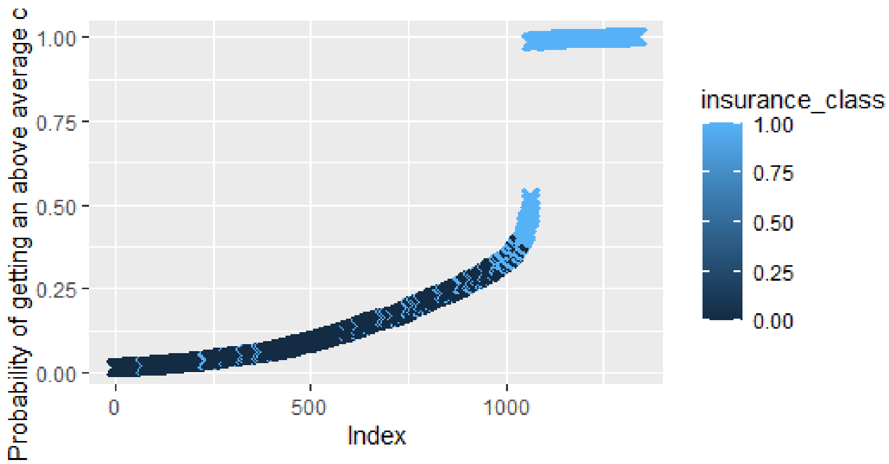

Figure 7: Probability of above average

This plot shows the insurance_class based on the ranking of probability of having or not to play an above average insurance_class. Since most of the low probability class didn't have to pay an above average charge whereas high probability ones did, shows that our model is performing quite well!