Inferential-Statistics-Report

October 18, 2017

1 Inferential Statistics Report

In this brief document, we summarize the inferential statistics analysis and finding from the Jupyter notebook `Inferential-Statistics.ipynb`.

1.1 Significance level (α)

In all of the hypothesis tests in our analysis, we used a significance level of $\alpha = 0.05$.

1.2 Our Questions

The questions that we tried to answer were as follows: 1. Is the proportion of defaults the same for men and women? 2. Is age a significant predictor of default? 3. Is credit limit a significant predictor of default? 4. Is the ratio of $\left(\frac{\text{bill amount}}{\text{credit limit}}\right)$ a significant predictor of default? Here, "bill amount" stands for past credit card bill amounts. 5. Is the ratio of $\left(\frac{\text{bill amount}-\text{pay amount}}{\text{credit limit}}\right)$ a significant predictor of default?

1.2.1 1. Is the proportion of defaults the same for men and women?

We wanted to test whether the proportion of defaults the same for men and women.

Let p_m represent the proportion of defaults for men.

Let p_w represent the proportion of defaults for women.

Our null and alternative hypotheses are as follows:

- H_0: $p_m = p_w$
- H_1: $p_m \neq p_w$

We used bootstrapping to test our null hypothesis. We reject the null hypothesis that $p_m = p_w$.

1.2.2 2. Is age a significant predictor of default?

We conducted a logistic regression where age was the predictor variable and default status was the target variable.

We used the implementation of logistic regression in the `glm` package in the R language. We chose to use `glm`'s implementation because it calculates the p-values associated with each regression coefficient. The logistic regression implemented in `scikit-learn` does not calculate these p-values.

We used the `rpy2` Python library to call R from within Python.

From our regression results, we found that the p-value for the regression coefficient was 0.0162, which is less than $\alpha = 0.05$. Therefore, we conclude that age was a statistically significant predictor of default.

The regression coefficient was positive, implying that the log-odds of default increase as age increases.

1.2.3 3. Is credit limit a significant predictor of default?

We conducted a logistic regression where credit limit is the predictor variable and default status is the target variable.

From our regression results, we found that the p-value for the regression coefficient was less than 2×10^{-16}, which is less than $\alpha = 0.05$. Therefore, we concluded that credit limit was a statistically significant predictor of default.

The regression coefficient was negative, implying that the log-odds of default decrease as credit limit increases.

1.2.4 4. Is the ratio of $\left(\frac{\textbf{bill amount}}{\textbf{credit limit}}\right)$ a significant predictor of default?

Here, "bill amount" stands for past credit card bill amounts.

We conducted a logistic regression where the ratio of $\left(\frac{\text{bill amount}}{\text{credit limit}}\right)$ was the predictor variable and default status was the target variable.

There are six bill amount features: 'BILL_AMT1', 'BILL_AMT2', ..., & 'BILL_AMT6'.

From our regression results, we found that the p-value for the regression coefficient of each bill amount ratio was less than 2×10^{-16}, which is less than $\alpha = 0.05$. Therefore, we concluded that, for each of the 6 bill amounts, the ratio of $\left(\frac{\text{bill amount}}{\text{credit limit}}\right)$ was a statistically significant predictor of default.

The regression coefficients were positive, implying that the log-odds of default increase as the ratio of $\left(\frac{\text{bill amount}}{\text{credit limit}}\right)$ increases.

1.2.5 5. Is the ratio of $\left(\frac{\textbf{bill amount}-\textbf{pay amount}}{\textbf{credit limit}}\right)$ a significant predictor of default?

We conducted a logistic regression where the ratio of $\left(\frac{\text{bill amount}-\text{pay amount}}{\text{credit limit}}\right)$ was the predictor variable and default status was the target variable.

From our regression results, we found that the p-value for the regression coefficient of each ratio is less than 2×10^{-16}, which is less than $\alpha = 0.05$. Therefore, we concluded that, for each of the 6 (bill amount, pay amount) pairs, the ratio of $\left(\frac{\text{bill amount}-\text{pay amount}}{\text{credit limit}}\right)$ was a statistically significant predictor of default.

The regression coefficients were positive, implying that the log-odds of default increase as the ratio of $\left(\frac{\text{bill amount}-\text{pay amount}}{\text{credit limit}}\right)$ increases.