ECO 520- Business Analytics Tool II

Assignment 3

Professor. Peter Bondarenko

Thuy Nguyen

**Question 1:**

**1.1**

A screenshot of a computer

Description automatically generated

**1.2**

**Multiple linear regression with Charges being the dependent variable:**

= -2,136.454 +266.544\* + 23,892.570\*

**1.3**

A screenshot of a paper

Description automatically generated

**1.4**

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 19 is 6,417.984

**1.5**

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 27 is 9,563.073

**1.6**

**Multiple linear regression with log(charges) being the dependent variable:**

= 7.299 +0.035\* + 1.546\*

**1.7**

**A white sheet with black text

Description automatically generated**

**1.8**

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 19 is 6,697.871

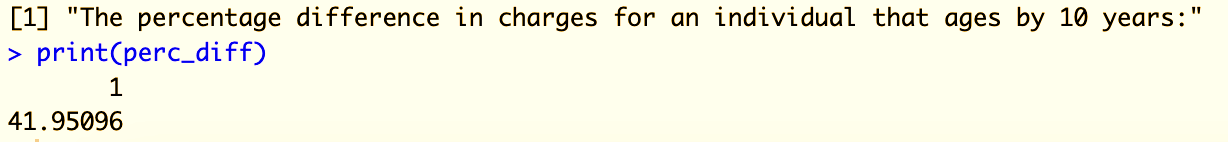
**1.9**

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 27 is 7,534.035

**1.10**

With age coefficient being 0.035, every 10 years of age results a change of 35% in log(charges).

In another word, there’s 41.9% in actual charges change for every 10-year increase in age.



**1.11**

= 4.16 + 1.264\* + 1.544\*

**1.12**

**A table of numbers and a list of numbers

Description automatically generated with medium confidence**

**1.13**

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 19 is 7,104.192

**1.14**

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 27 is 8,029.289

**1.15**

The coefficient β1=1.264 indicates that a **1% increase in age** is associated with approximately a **1.264% increase in charges** due to the use of logarithm for both charges and age.

**Question 2:**

**2.1  
A screenshot of a computer

Description automatically generated**

**2.2**

**Checking account statistical summary:**

Min: -665

1st Quartile: 164.8

Median: 351.5

Mean: 362.4

3rd Quartile: 553.5

Max: 1319.0

**Term statistical summary:**

Min: 9

1st Quartile: 16

Median: 18

Mean: 17.82

3rd Quartile: 20

Max: 27

**Credit score statistical summary:**

Min: 376

1st Quartile: 725.8

Median: 770.5

Mean: 760.5

3rd Quartile: 812

Max: 1029

**Amount statistical summary:**

Min: 244

1st Quartile: 1016

Median: 1226

Mean: 1219

3rd Quartile: 1420

Max: 2362

**Saving amount statistical summary:**

Min: 2082

1st Quartile: 2951

Median: 3203

Mean: 3179

3rd Quartile: 3402

Max: 4108

**Employment duration statistical summary:**

Min: 0

1st Quartile: 15

Median: 41

Mean: 49.39

3rd Quartile: 85

Max: 120

**Age statistical summary:**

Min: 18

1st Quartile: 29

Median: 32

Mean: 31.21

3rd Quartile: 34

Max: 42

**Number of credit account statistical summary:**

Min: 1

1st Quartile: 1

Median: 2

Mean: 2.546

3rd Quartile: 3

Max: 9

**2.3**

**Default:** 300 defaulted; 700 not defaulted

**Gender:** 310 female; 690 male

**Marital status:** 548 married; 452 single

**Car loan:** #0: 647; #1: 353 (1 if person has a car loan, 0 otherwise)

**Personal loan:** #0: 526; #1: 474 (1 if person has a personal loan, 0 otherwise)

**Home loan:** #0: 944; #1: 56 (1 if person has a home loan, 0 otherwise)

**Education loan:** #0: 888; #1: 112 (1 if person has a student loan, 0 otherwise)

**Employment status:** 304 employed and 692 unemployed

**2.4**

Yes, there seems to be enough variation in the categorical variables to build a reliable model for loan defaults.

**2.5**

**Default =** 3.465- 0.0003\*Checking Amount + 0.014\*Term – 0.001\*Credit Score + 0.008\*Gender (Male) – 0.042\*Marital Status (Single)- 0.079\*Car Loan – 0.147\*Personal Loan – 0.215\*Home Loan + 0.043\*Education Loan + 0.052\* Employment Status (Unemployed)+ 0.0001\*Amount – 0.00003\*Saving Amount -0.0002 \* Employment Duration – 0.044\*Age – 0.01\*Number of credit accounts

**2.6**

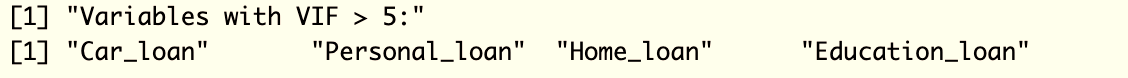
A paper with numbers and text

Description automatically generated

**2.7**

**A graph of values for independent variables

Description automatically generated**



By using VIF, the following independent variables exhibit a high degree of multicollinearity: Car loan, Personal loan, Home loan, Education loan

**2.8**

**Default =** 3.472 -0.0003\*Checking amount + 0.016\*Term -0.001\*Credit score +0.028\*Employment status (Unemployed) + 0.0001\*Amount -0.0003\*Saving account -0.047\*Age

**2.9**

**A paper with numbers and text

Description automatically generated**

**2.10**

The multiple linear regression model can’t be used as a model of loan default probabilities because the dependent variable should’ve fallen into the valid range of probabilities, which is [0,1]. A probability greater than 1 or less than 0 is not valid. Below is the result for example test.



**Question 3:**

**3.1**

**=** 38.294 -0.005\*Checking amount + 0.178\*Term -0.012\*Credit score +0.485\*Employment status (Unemployed) +0.0005\*Amount -0.0005\*Saving amount -0.626\*Age

**3.2**

**A paper with numbers and text

Description automatically generated**

**3.3**

**=** 38.848 -0.005\*Checking amount + 0.175\*Term -0.011\*Credit score -0.005\*Saving amount -0.629\*Age

**3.4**

**A table of numbers and text

Description automatically generated with medium confidence**

**3.5**

With employment status being insignificant, it doesn’t affect the loan default probabilities if an individual is employed or not.

**3.6**

Running an example dataset for 2 similar individuals with $500 in checking, term is 22, (credit) amount is $2000, saving amount is $2200 and age is 35, the prediction returns:

For an individual with a 600-credit score, their loan default probability is 81.89% (Very likely to default the loan) while a similar individual with a 800-credit score has 31.66% loan default (<50%, not likely to default the loan). The percentage difference for probability prediction is around 50.24%

A close-up of a computer code

Description automatically generated

**Question 4:**

**4.1**



**4.2**

**=** 36.714 -0.004\*Checking amount + 0.157\*Term -0.012\*Credit score -0.004\*Saving amount -0.63\*Age

**4.3**

**A paper with numbers and symbols

Description automatically generated**

**4.4**

Preview of the probabilities when running training logistic model on testing dataset, all are within the [0,1] range: **A table of numbers with numbers

Description automatically generated**

**4.5**

Classification table:

A screenshot of a computer

Description automatically generated

**4.6**

The accuracy rate of my model as the ratio of correctly predicted outcomes over the total possible outcomes is 91.33%

A close up of a word

Description automatically generated

**Question 5:**

**5.1**

- The model relied on concentration points where most paying Prime members live however this approach ignored the structural inequalities. Communities other than white tend to have lower Prime rates and that led to the exclusion from new Prime same-day delivery service.

-The model did not consider about people that live in those minority neighborhoods have less access to transportation or other quality retailer stores.

- Even though the leader claimed that this is not race-based, the inequality reflected that a lot of areas in big cities in NYC, Chicago, Atlanta, etc. were left out, especially when those areas usually face higher cost of living.

- Paying customers that pay the same annual Prime memberships did not expect to not receive the same service that someone live 15 minutes away from them do. There’s lack of transparency of what indicates eligibility for each area even though research showed crime rates, income inequality somewhat corresponded with the service’s exclusion.

**5.2**

Some problems associated with using ML algorithms to assist with decision making in the justice system are:

* Discrimination: ML uses historical data, which would show inequalities and racial bias in the past.
* Lack of transparency: Nowadays, transparency is very important however, a decision is made based on a lot of people’s work and it’s hard to tell whose responsibility for errors.
* Ethnical concerns: It’s not appropriate to use algorithms to make decisions that could affect someone’s freedom or livelihood.