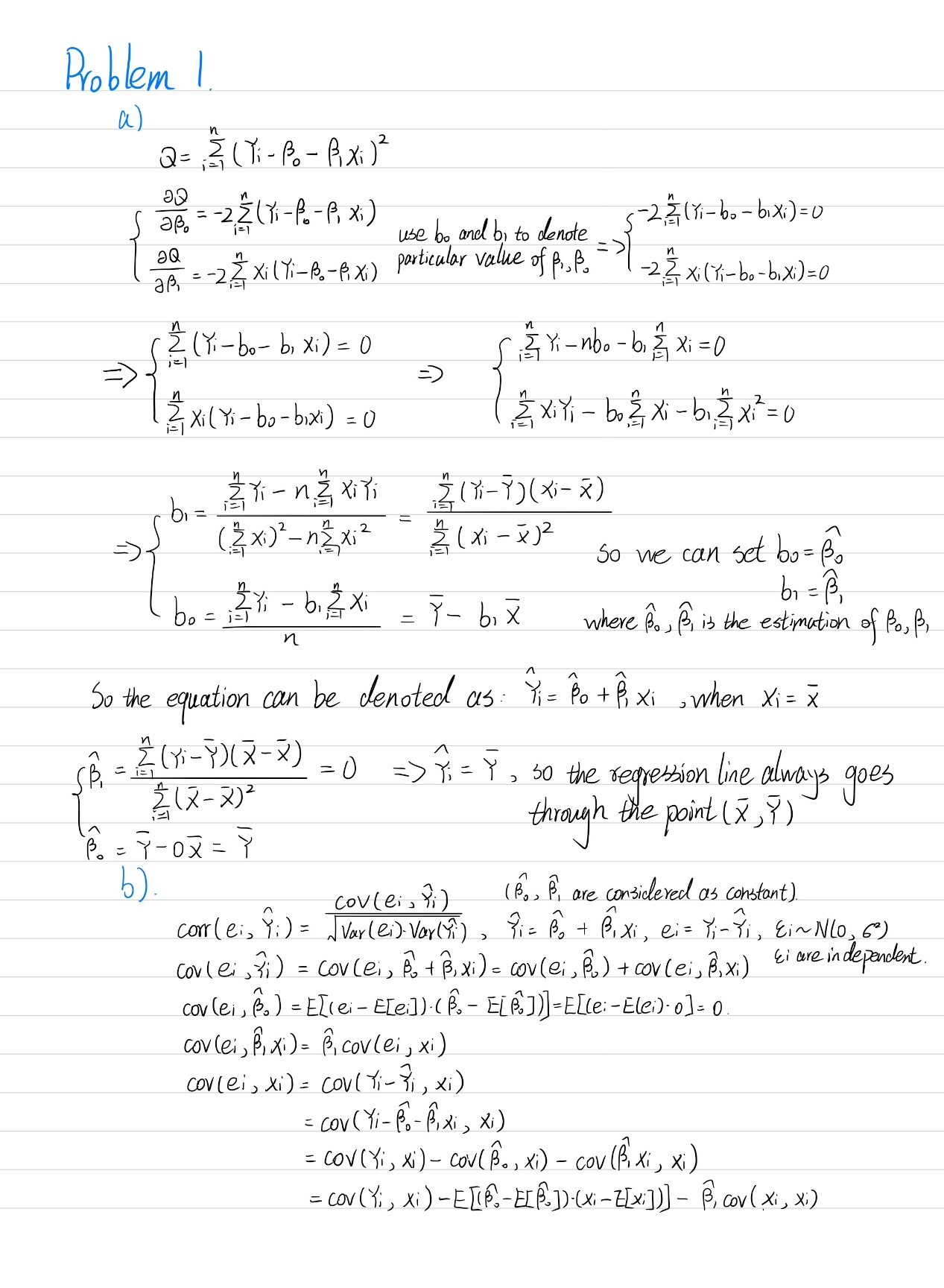
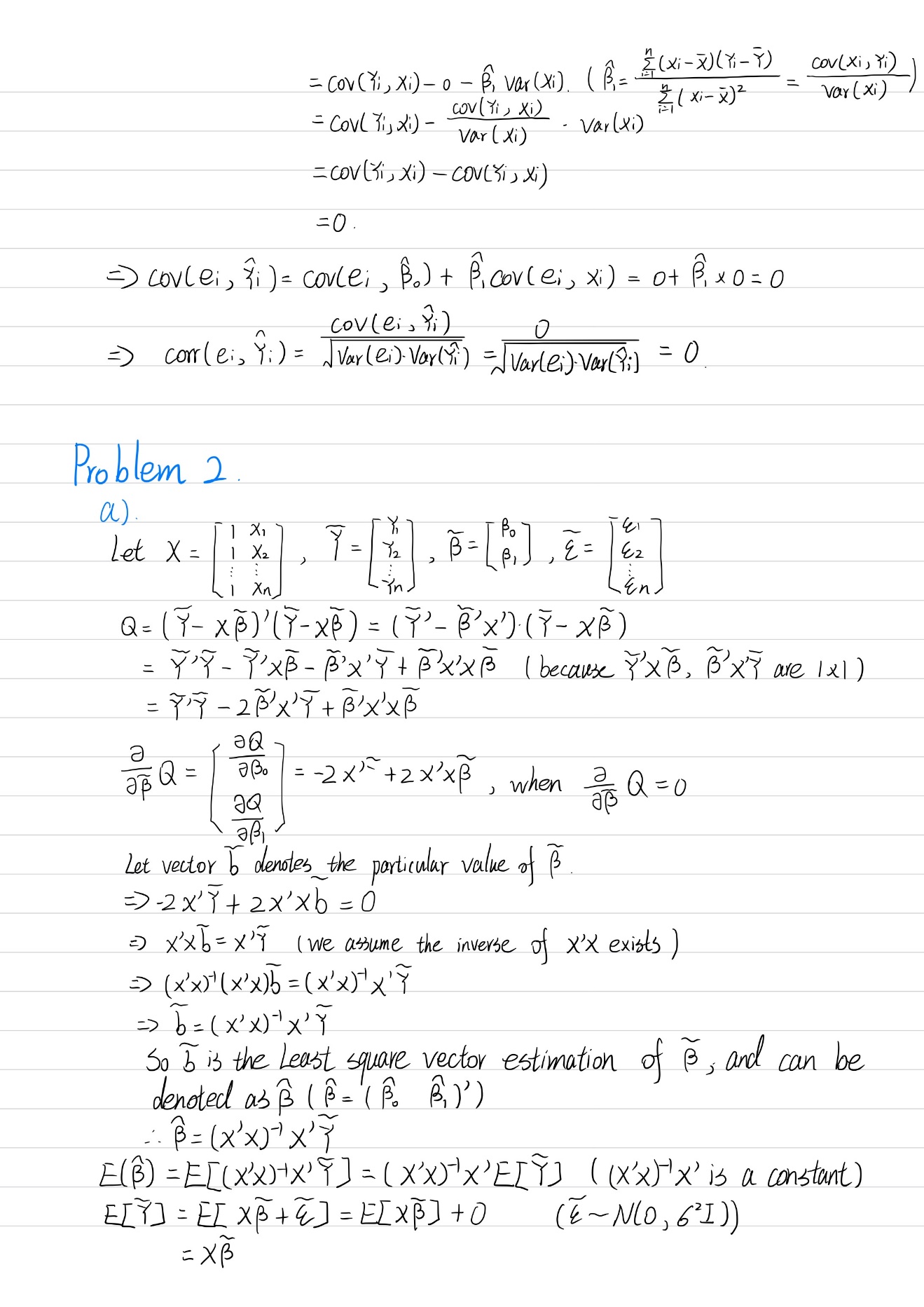
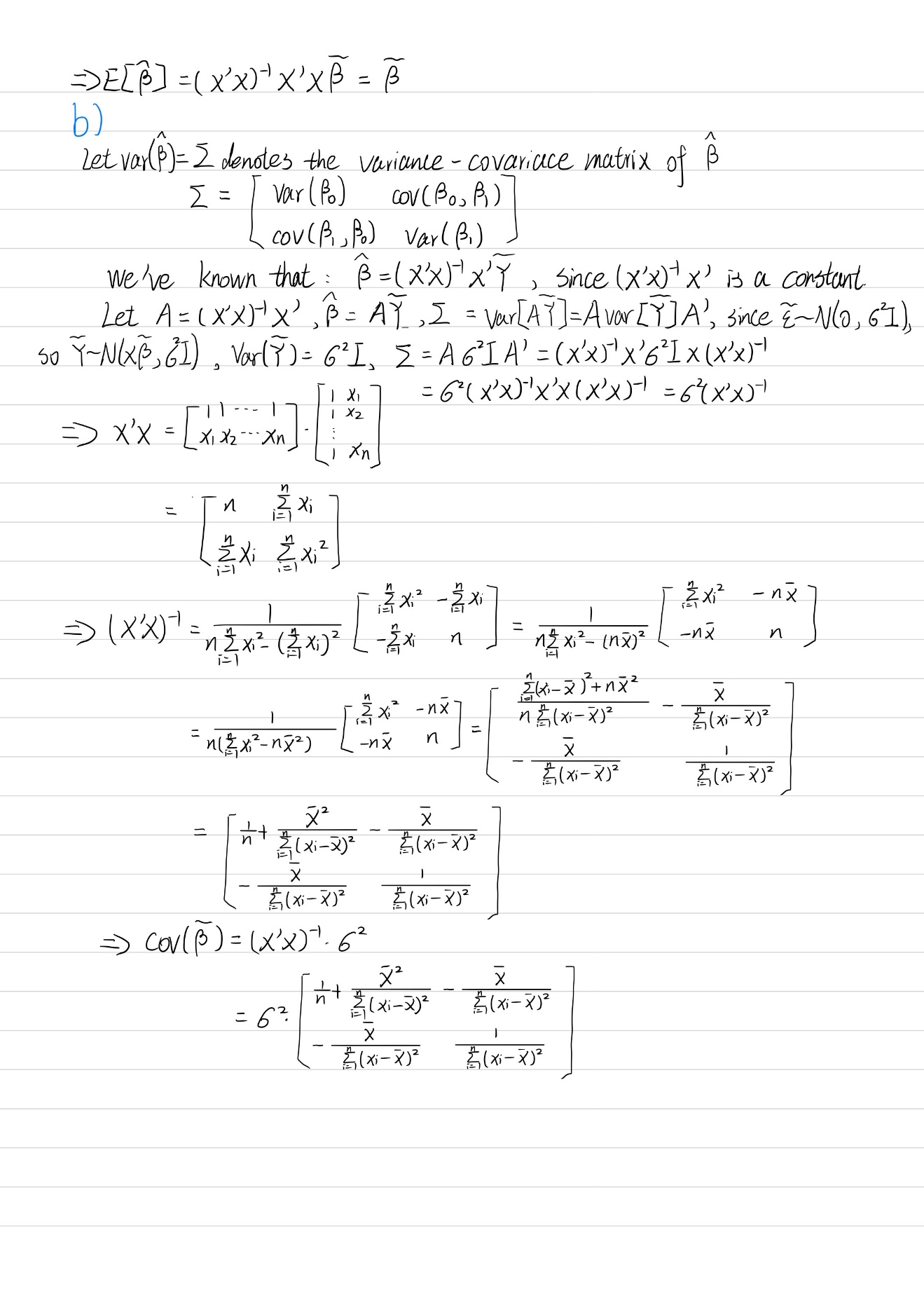
p8130\_hw4\_rq2166

Ruoyuan Qian







# Problem 3

## a)

The linear regression can be denoted as follow:

, with 

In this situation, glia-neuron ratio denotes the *Y*, *ln* (brain mass) denotes the *X*. Least squares estimation is implemented to estimate the coefficient.









Let , denote the estimation of true coefficient , , so we obtain:







It can be calculated in R software, we can obtain:

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 0.16370 0.15987 1.024 0.322093

## ln\_brain\_mass 0.18113 0.03604 5.026 0.000151 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.1699 on 15 degrees of freedom

## Multiple R-squared: 0.6274, Adjusted R-squared: 0.6025

## F-statistic: 25.26 on 1 and 15 DF, p-value: 0.0001507

60% of the variance of Y (glia-neuron ratio) explained by X (*ln* (brain mass)).

## b)

The *ln*(brain mass) of human is equal to 7.22, let denote the mean fitted outcome with human glia-neuron ratio, so we can estimate through the regression above:



## c)

An interval for the prediction of a single new observation is more relevant. Since the nonhuman data is used to build the regression line, human observation is considered as a new independent trial.

## d)

Since we have calculated the estimated value of glia-neuron ratio for humans, , the prediction interval (PI) can be calculated as follow:



In R we can obtain prediction interval by:

predict.lm(fit, interval="prediction")

## fit lwr upr

## 1 1.471458 1.036047 1.906869

From the output of linear regression, 95% prediction interval for  is (1.04, 1.91). We are 95% confident that the value of glia-neuron ratio for humans will be between 1.04 and 1.91.

No, the upper bound of prediction of glia-neuron ratio for human is 1.91, and the observed glia-neuron ratio for human is 1.65, the observed value falls into the prediction interval, which means the human brain do not have excessive glia-neuron ratio for its mass compared with other primates.

## e)

We need to worry about extrapolation, since the dataset which the model is based on of the *ln*(brain mass) is between about 2 and 6. However, the *ln*(brain mass) for human is 7.22, which is totally out of the boundary, so the prediction of human brain is not reliable.

# Problem 4

## a)

**Description**:

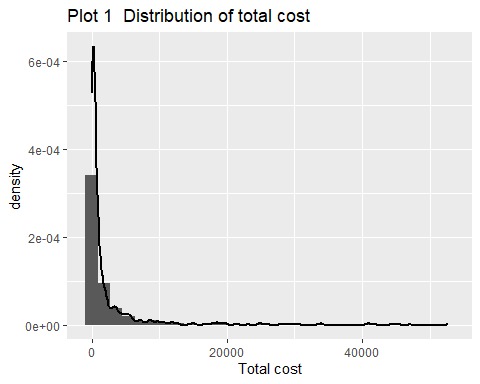
The dataset is collected by an insurance company about the total cost of heart disease and other relevant events caused by heart disease. The main outcome is total cost from the heart disease, and the main predictor is the number of emergency room (ER) visits. There are other important variables like age and gender of subject, the number of other complications arose due to other disease instead of heart disease, the number of days of treated duration.

Table 1 Descriptive statistics for variable of interest

|  |  |
| --- | --- |
| **totalcost** | |
| Mean (SD) | 2799.956 (6690.260) |
| Median (Q1, Q3) | 507.200 (161.125, 1905.450) |
| Range | 0.000 - 52664.900 |
| **age** | |
| Mean (SD) | 58.718 (6.754) |
| Median (Q1, Q3) | 60.000 (55.000, 64.000) |
| Range | 24.000 - 70.000 |
| **gender** | |
| female | 608 (77.2%) |
| male | 180 (22.8%) |
| **interventions** | |
| Mean (SD) | 4.707 (5.595) |
| Median (Q1, Q3) | 3.000 (1.000, 6.000) |
| Range | 0.000 - 47.000 |
| **drugs** | |
| Mean (SD) | 0.447 (1.064) |
| Median (Q1, Q3) | 0.000 (0.000, 0.000) |
| Range | 0.000 - 9.000 |
| **e\_rvisits** | |
| Mean (SD) | 3.425 (2.637) |
| Median (Q1, Q3) | 3.000 (2.000, 5.000) |
| Range | 0.000 - 20.000 |
| **complications** | |
| Mean (SD) | 0.057 (0.248) |
| Median (Q1, Q3) | 0.000 (0.000, 0.000) |
| Range | 0.000 - 3.000 |
| **comorbidities** | |
| Mean (SD) | 3.766 (5.951) |
| Median (Q1, Q3) | 1.000 (0.000, 5.000) |
| Range | 0.000 - 60.000 |
| **duration** | |
| Mean (SD) | 164.030 (120.916) |
| Median (Q1, Q3) | 165.500 (41.750, 281.000) |
| Range | 0.000 - 372.000 |
| **Overall(N=788)** | |

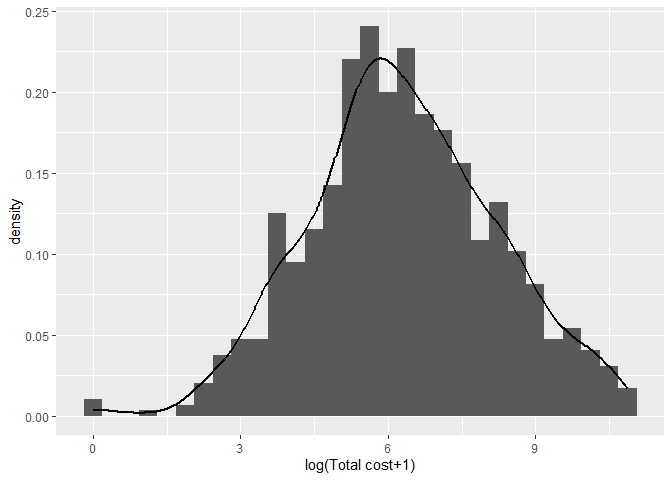
## b)

According to Plot 1, the distribution of total cost is extremely skewed, in order to fit the linear regression, transformation is needed.



Log-transformation is done, however, 3 observations which are equal to 0 will become negative infinite due to the *ln*(totalcost) transformation, so *ln*(totalcost+1) is implemented to avoid infinite value.

According to the Plot 2, the distribution after log-transformation is quite normal, so the transformed data can be implemented into the linear regression model.



Plot 2 The Distributions and Q-Q plot for total cost after log-transformation

## c)

New variable is created through R software:

table(heart\_log\_bin$comp\_bin)/length(heart\_log\_bin$comp\_bin)

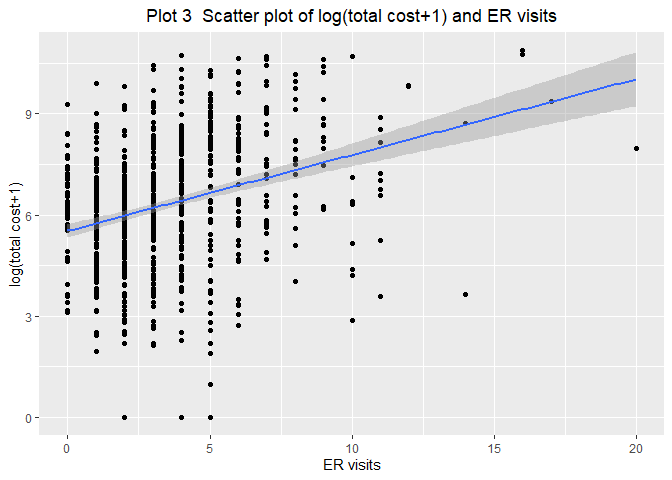
## 0 1

## 0.94543147 0.05456853

The proportion of 0 in the new variable is 94.54%, and the proportion of 1 is 5.46%.

## d)

Scatter plot is drawn between *ln* (total cost+1) and ER visits as follow:



Plot 3 Scatter plot of log(total cost+1) and ER visits

The linear regression can be denoted as follow:

, with .

In this situation, *ln* (total cost+1) denotes the *Y*, ER visits denotes the *X*. Least squares estimation is implemented to estimate the coefficient.

It can be calculated in R software, we can obtain:

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 5.52674 0.10510 52.584 <2e-16 \*\*\*

## e\_rvisits 0.22529 0.02432 9.264 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.799 on 786 degrees of freedom

## Multiple R-squared: 0.09844, Adjusted R-squared: 0.09729

## F-statistic: 85.82 on 1 and 786 DF, p-value: < 2.2e-16

Let , denote the estimation of true coefficient , , so we obtain:







The slope is tested to determine whether there is a significant relation between X and Y. A two-sided one sample t-test is used.

Hypothesis:

, where 

Significant level:



The test-statistics:

,

,

,



At significance level of 0.05, , we reject the null and conclude that there is sufficient evidence that coefficient of ER visits is not equal to zero, which means ER visits is a significant predictor of *ln* (total cost+1).

Interpretation of the slope: For every unit increase in the number of emergency room visits, the estimated *ln* (total cost+1) increases by 0.22529 dollars, on average.

## e)

### i)

Multiple linear regression (MLR) with ‘comp\_bin’ and ‘ERvisits’ as predictors can be fitted:



It can be calculated in R software, we can obtain:

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 5.48849 0.10500 52.271 < 2e-16 \*\*\*

## e\_rvisits 0.20947 0.02490 8.412 < 2e-16 \*\*\*

## factor(comp\_bin)1 2.19096 0.55447 3.951 8.47e-05 \*\*\*

## e\_rvisits:factor(comp\_bin)1 -0.09753 0.09630 -1.013 0.311

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.759 on 784 degrees of freedom

## Multiple R-squared: 0.1405, Adjusted R-squared: 0.1372

## F-statistic: 42.72 on 3 and 784 DF, p-value: < 2.2e-16

The objective of testing if comp\_bin is an effect modifier of the relationship between total cost and ERvisits can be transformed to test whether the coefficient of product of comp\_bin and ERvisits () is significantly unequal to zero.

So the hypothesis of the test can be written as:

 where 

Significant level:



The test-statistics:

,

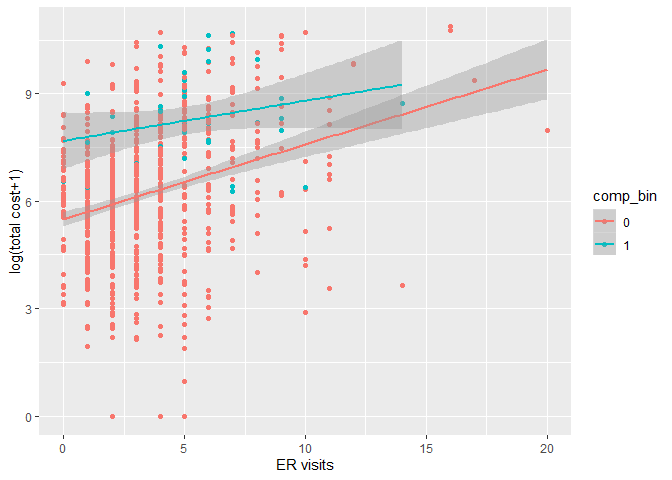
,

,



At significance level of 0.05, , we fail to reject the null and conclude that there is sufficient evidence that coefficient of product of comp\_bin and ERvisits is equal to zero, which means comp\_bin is not an effect modifier of the relationship between total cost and ERvisits.

Also, we can make plot to compare the slopes grouped by comp\_bin, the slopes are closely parallel, which means it does not an effect modifier as well.



Plot 4 Regression lines grouped by comp\_bin

### ii)

The objective of testing if comp\_bin is a confounder of the relationship between total cost and ERvisits can be transformed to test whether the changing rate of coefficient of ERvisits () is smaller than 10%.

Two regression models can be built for the test:

Regression 1: 

Regression 2: 

It can be calculated in R software, we can obtain:

Regression 1:

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 5.52674 0.10510 52.584 <2e-16 \*\*\*

## e\_rvisits 0.22529 0.02432 9.264 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.799 on 786 degrees of freedom

## Multiple R-squared: 0.09844, Adjusted R-squared: 0.09729

## F-statistic: 85.82 on 1 and 786 DF, p-value: < 2.2e-16

Regression 2:

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 5.51020 0.10279 53.606 < 2e-16 \*\*\*

## e\_rvisits 0.20295 0.02405 8.437 < 2e-16 \*\*\*

## factor(comp\_bin)1 1.70573 0.27915 6.111 1.56e-09 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.759 on 785 degrees of freedom

## Multiple R-squared: 0.1394, Adjusted R-squared: 0.1372

## F-statistic: 63.57 on 2 and 785 DF, p-value: < 2.2e-16

Changing rate:



Since the changing rate of coefficient of ERvisits is almost equal to 10%, the comp\_bin is a confounder between *ln*(total cost+1) and ERvisits.

### iii)

Two regressions can be tested through ANOVA:

Regression 1: 

Regression 2: 

It can be calculated in R software, we can obtain:

## Analysis of Variance Table

##

## Model 1: log\_totalcost ~ e\_rvisits

## Model 2: log\_totalcost ~ e\_rvisits + factor(comp\_bin)

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 786 2544.8

## 2 785 2429.3 1 115.55 37.339 1.563e-09 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Hypothesis:

*H0:* ()*,*

*H1:* ()

Significant level:



The test-statistics:

,

,

,



At significance level of 0.05, , we reject the null and conclude that there is sufficient evidence that model with comp\_bin is superior, so it is defensible to add comp\_bin variable into the model.

## f)

### i)

The linear regression can be denoted as follow:



It can be calculated in R software, we can obtain:

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 5.9404610 0.5104064 11.639 < 2e-16 \*\*\*

## e\_rvisits 0.1745975 0.0225736 7.735 3.20e-14 \*\*\*

## factor(comp\_bin)1 1.5044946 0.2584882 5.820 8.57e-09 \*\*\*

## age -0.0206475 0.0086746 -2.380 0.0175 \*

## duration 0.0057150 0.0004888 11.691 < 2e-16 \*\*\*

## gendermale -0.2067662 0.1387002 -1.491 0.1364

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.624 on 782 degrees of freedom

## Multiple R-squared: 0.2694, Adjusted R-squared: 0.2647

## F-statistic: 57.68 on 5 and 782 DF, p-value: < 2.2e-16



For one unit (one time) increases in emergency room visit, the estimated *ln*(total cost+1) will increase by 0.175 dollars on average, adjusted for age, number of days of duration of treatment conditions, gender and whether there is/are complication(s).

26.47% of the variance of Y (*ln*(total cost+1)) can be explained by X (ERvisits).

### ii)

Two regressions can be tested through ANOVA:

It can be calculated in R software, we can obtain:

## Analysis of Variance Table

##

## Model 1: log\_totalcost ~ e\_rvisits

## Model 2: log\_totalcost ~ e\_rvisits + factor(comp\_bin) + age + duration +

## gender

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 786 2544.8

## 2 782 2062.2 4 482.62 45.753 < 2.2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Hypothesis:

*H0:* ()*, H1:* ororor

()

Significant level:



The test-statistics:

,

,

,



At significance level of 0.05, , we reject the null and conclude that there is sufficient evidence that at least one of the coefficients () is not equal to zero, so the large model is superior. It is defensible to add comp\_bin, age, gender and duration into the model.

I will use the MLR to address the interest problem of investigator. In the MLR, the coefficient of main predictor (ERvisits) is more precise due to the adjustment of other covariates, which may bias the result. Furthermore, the adjust R square is much larger in the large model (0.2647)than in the small model (0.09729).