Final Group Report MATH4322

University of Houston

MATH4322

ECONOMIC EFFECTS OF HEALTHCARE FOR VIETNAMESE CITIZENS

Riches Dang

Chris Do

Justin George

John Pham

Introduction (John Pham)

Healthcare has been a necessity to humans since the creation of communities. We, in the modern era, are fortunate to be able to have access to healthcare that populations before us have not been able to have. However, although we have this privilege, there is still one major issue that arises – cost. Money and economic prosperity are benefits not every person has in the world. Americans know this very well, as they do not have access to free healthcare. We wish to gain insight on what the medical expenditures are like for families of better opportunities. Vietnam is one of these countries, hence why we have decided to choose it as our topic to delve into.

The Vietnam World Bank Livings Standards Survey has provided a dataset that includes 27,765 observations with 12 variables regarding the medical expenses in Vietnam at the individual level. We can utilize this data set by using the “Ecdat” package and “VietNamI” item in R. Our response variable will be ***lnhhexp***, a quantitative variable describing the logarithm of total amount of money spent on medical expenses.

Our goal is to answer the question: What is the economic analysis of healthcare expenditure for medical visits in the context of the Vietnamese healthcare system, accounting for variations in cost factors such as geographic location, patient demographics, and specific service components?

Linear Regression Model (Chris Do, John Pham)

With the response variable being ***lnhhexp***(quantitative), the goal is to predict the total amount of money spent on medical expenses given the predictors. A Linear Regression model would be the most forward approach for this data. An advantage of using this model is that it is simple and easy to interpret. The disadvantage of this model is that it assumes a linear relationship between the response variable and the predictors. Additionally, outliers can influence the model parameters and predictions disproportionately.

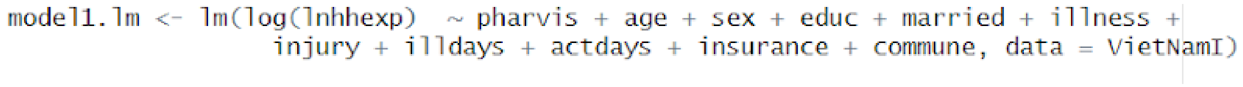
The model equation that we will be using for Linear Regression is the following:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \ldots + \varepsilon$$

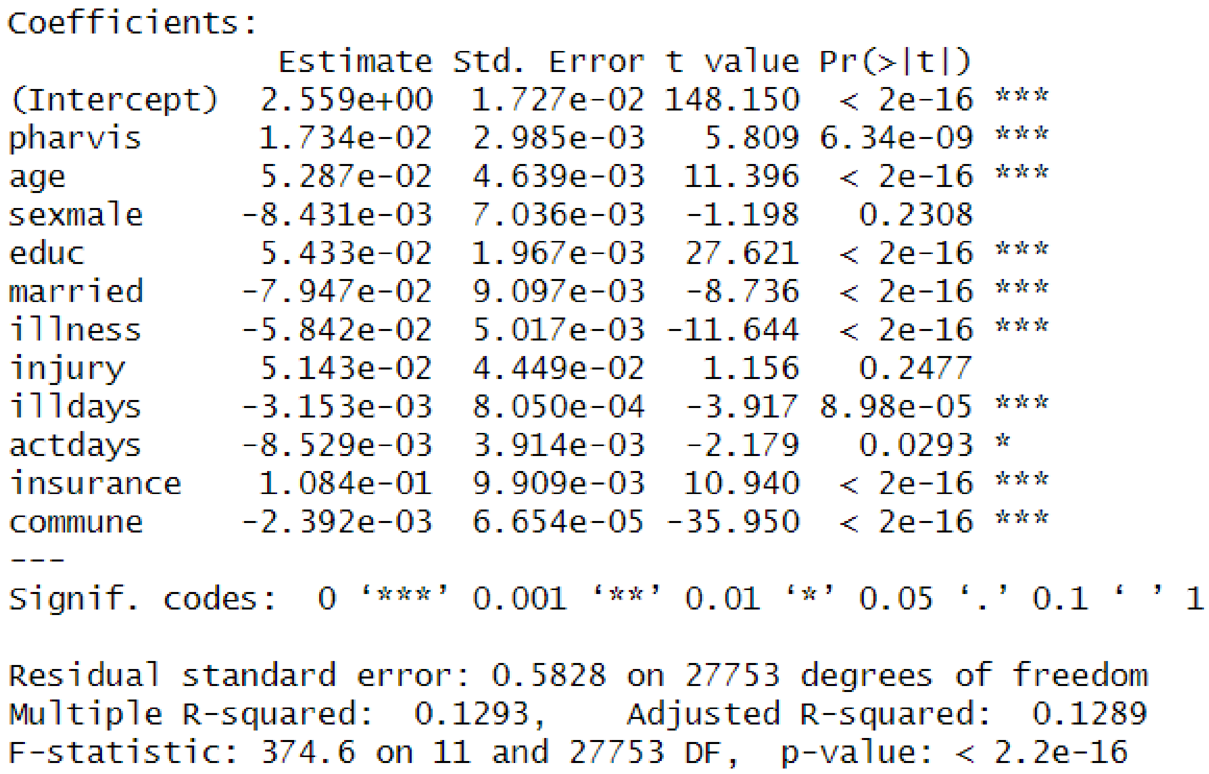
For linear regression, we will be excluding the predictors ***sexmale*** (gender of individual), ***injury*** (whether the individual was injured or not during the survey) from consideration because they are not significant regarding the response variable. Following the model represents the predictors **pharvis**, **age**, **educ**, **married**, **illness**, **illdays**, **actdays**, **insurance**, **commune,** with **β_i** as the repective coefficients.

Using the lm() function, we started our first regression model (model1.lm) by fitting all the predictors against the response variable ***lnhhexp**.*

Call:



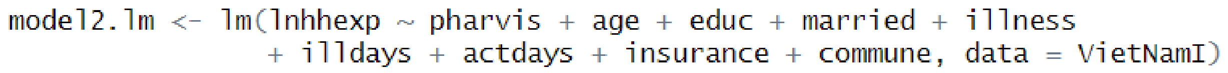
```
model1.lm <- lm(log(lnhhexp)  ~ pharvis + age + sex + educ + married + illness +
                    injury + illdays + actdays + insurance + commune, data = VietNamI)
```



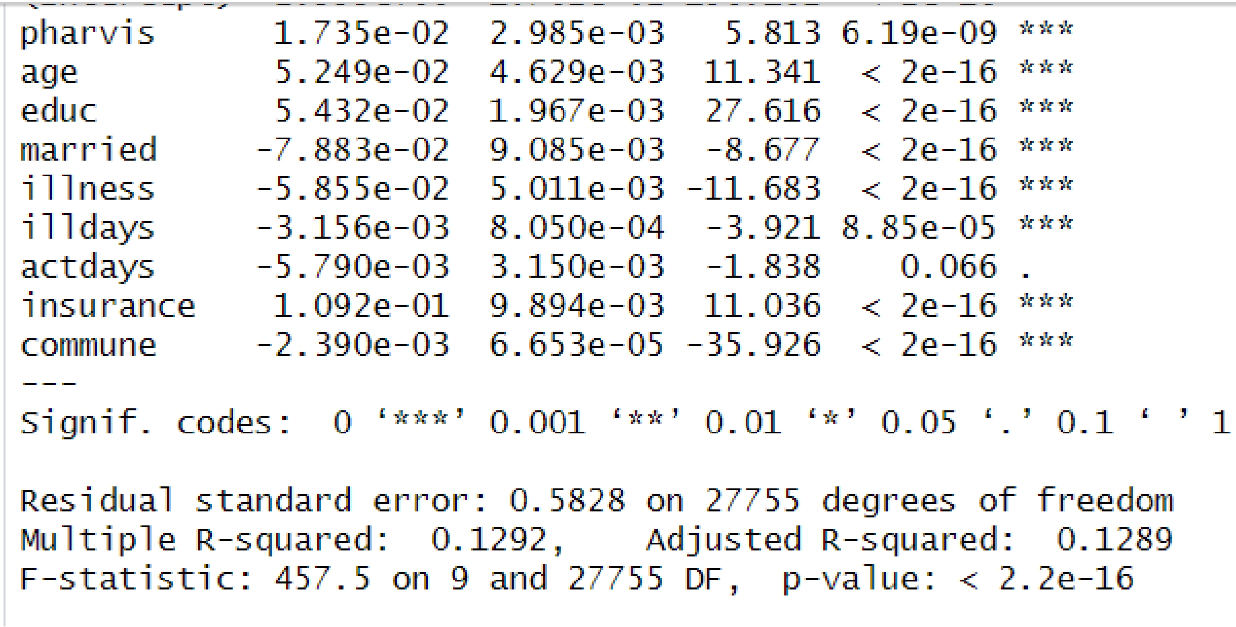
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.559e+00  1.727e-02 148.150  < 2e-16 ***
pharvis      1.734e-02  2.985e-03   5.809 6.34e-09 ***
age          5.287e-02  4.639e-03  11.396  < 2e-16 ***
sexmale     -8.431e-03  7.036e-03  -1.198   0.2308    
educ         5.433e-02  1.967e-03  27.621  < 2e-16 ***
married     -7.947e-02  9.097e-03  -8.736  < 2e-16 ***
illness     -5.842e-02  5.017e-03 -11.644  < 2e-16 ***
injury       5.143e-02  4.449e-02   1.156   0.2477    
illdays     -3.153e-03  8.050e-04  -3.917 8.98e-05 ***
actdays     -8.529e-03  3.914e-03  -2.179   0.0293 *  
insurance    1.084e-01  9.909e-03  10.940  < 2e-16 ***
commune     -2.392e-03  6.654e-05 -35.950  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5828 on 27753 degrees of freedom
Multiple R-squared:  0.1293,	Adjusted R-squared:  0.1289 
F-statistic: 374.6 on 11 and 27753 DF,  p-value: < 2.2e-16
```

From there we were able to exclude the predictors **sexmale** and **injury** due to their *t-statistic P-values* being greater than 0.05. After refitting the remaining predictors to model2.lm, actdays had a *t-statistic p-value* of 0.066 which is greater that 0.05 so we were able to exclude it for the next model.

Call:



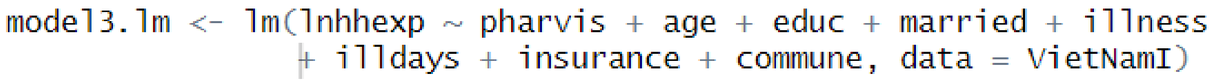
```
model2.lm <- lm(lnhhexp ~ pharvis + age + educ + married + illness
                + illdays + actdays + insurance + commune, data = VietNamI)
```



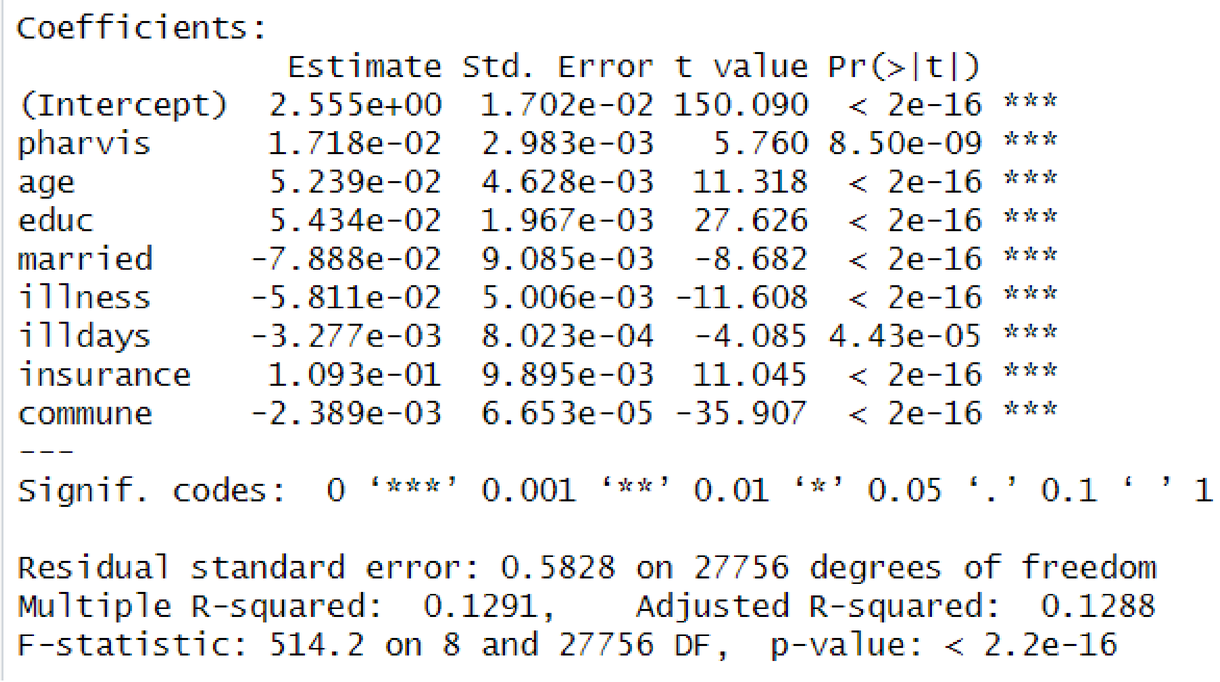
```
pharvis       1.735e-02  2.985e-03   5.813 6.19e-09 ***
age           5.249e-02  4.629e-03  11.341  < 2e-16 ***
educ          5.432e-02  1.967e-03  27.616  < 2e-16 ***
married      -7.883e-02  9.085e-03  -8.677  < 2e-16 ***
illness      -5.855e-02  5.011e-03 -11.683  < 2e-16 ***
illdays      -3.156e-03  8.050e-04  -3.921 8.85e-05 ***
actdays      -5.790e-03  3.150e-03  -1.838    0.066 .
insurance     1.092e-01  9.894e-03  11.036  < 2e-16 ***
commune      -2.390e-03  6.653e-05 -35.926  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5828 on 27755 degrees of freedom
Multiple R-squared:  0.1292,	Adjusted R-squared:  0.1289 
F-statistic: 457.5 on 9 and 27755 DF,  p-value: < 2.2e-16
```

Finally, model3.lm shows that the remaining predictors of **pharvis, age, educ, married, illness, actdays, illdays, insurance, commune** are significant to the model.

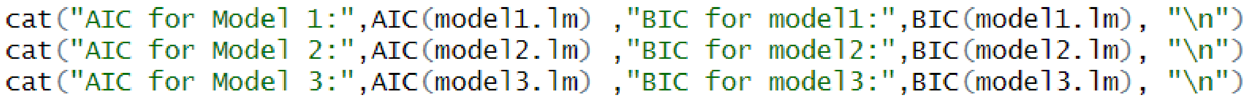
Call:



```
model3.lm <- lm(lnhhexp ~ pharvis + age + educ + married + illness
                + illdays + insurance + commune, data = VietNamI)
```



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.555e+00  1.702e-02 150.090  < 2e-16 ***
pharvis      1.718e-02  2.983e-03   5.760 8.50e-09 ***
age          5.239e-02  4.628e-03  11.318  < 2e-16 ***
educ         5.434e-02  1.967e-03  27.626  < 2e-16 ***
married     -7.888e-02  9.085e-03  -8.682  < 2e-16 ***
illness     -5.811e-02  5.006e-03 -11.608  < 2e-16 ***
illdays     -3.277e-03  8.023e-04  -4.085 4.43e-05 ***
insurance    1.093e-01  9.895e-03  11.045  < 2e-16 ***
commune     -2.389e-03  6.653e-05 -35.907  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5828 on 27756 degrees of freedom
Multiple R-squared:  0.1291,	Adjusted R-squared:  0.1288 
F-statistic: 514.2 on 8 and 27756 DF,  p-value: < 2.2e-16
```



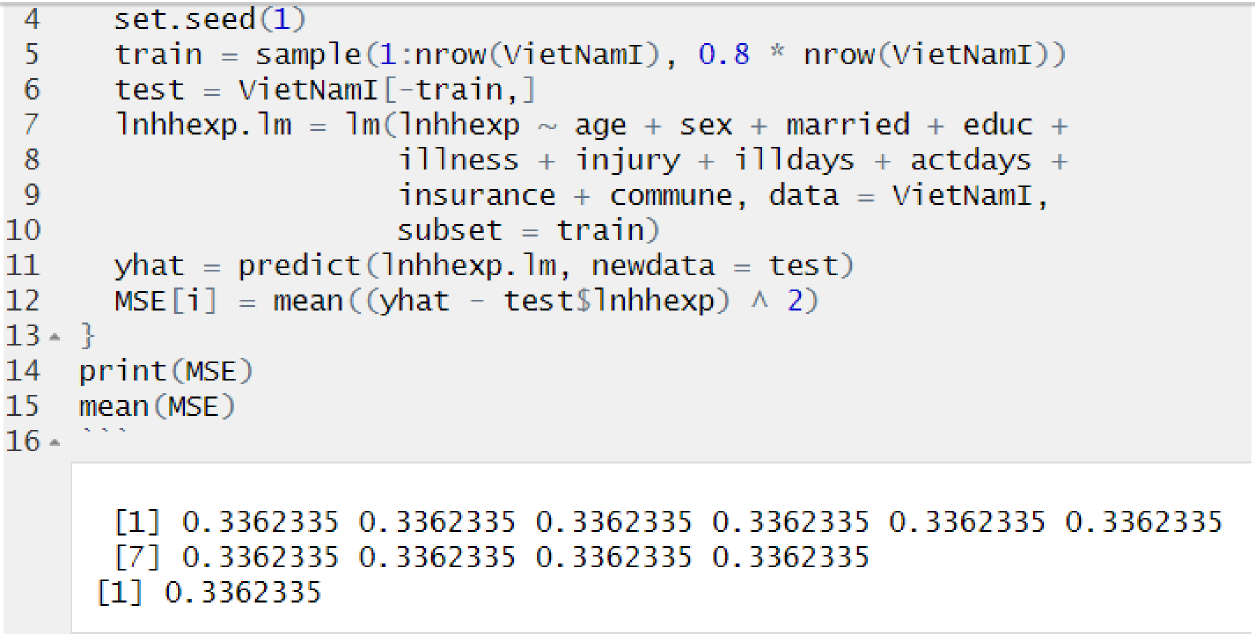
```
cat("AIC for Model 1:",AIC(model1.lm) ,"BIC for model1:",BIC(model1.lm), "\n")
cat("AIC for Model 2:",AIC(model2.lm) ,"BIC for model2:",BIC(model2.lm), "\n")
cat("AIC for Model 3:",AIC(model3.lm) ,"BIC for model3:",BIC(model3.lm), "\n")
```

```
AIC for Model 1: 48824.05 BIC for model1: 48931.06
Adjusted RSq for model 1: 0.1289229

AIC for Model 2: 48822.87 BIC for model2: 48913.42
Adjusted RSq for model 2: 0.128897

AIC for Model 3: 48824.25 BIC for model3: 48906.57
Adjusted RSq for model 3: 0.1288224
```

The formula for the new equation can be represented by

lnhhexp = 2.555 + 0.01718 x **pharvis** + 0.05239 x **age** + 0.05434 x **educ**
- 0.07888 x **married** - 0.05811 x **illness** - 0.003277 x **illdays**
+ 0.1093 x **insurance** -0.002389 x **commune**

Based on this model, some of the predictors will have a positive effect (pharvis, age, educ, insurance) while the rest will have a negative effect (married, illness, illdays). Holding the other predictors constant, for every one unit of increase in **pharvis**, **lnhhexp** will increase by 0.01718; for every unit of increase in **age**, **lnhhexp** will increase by 0.05239; for every unit of increase in **educ**, **lnhhexp** will increase by 0.05434; for every unit increase in **insurance**, Inhhexp will increase 0.1093;for every unit increase in **married**, Inhhexp will decrease by 0.07888; for every unit increase in **illness**, Inhhexp will decrease by 0.05811; for every unit increase in illdays, Inhhexp will decrease by 0.003277;for every unit increase in commune, Inhhexp will decrease by 0.002389.



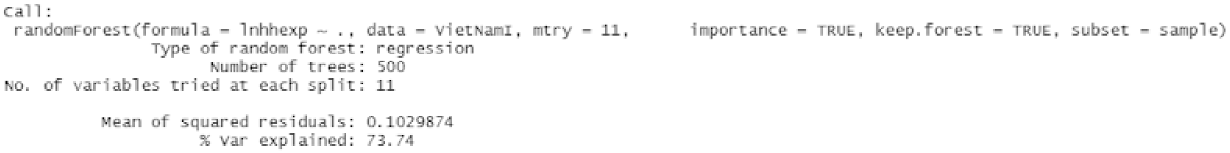
```
  set.seed(1)
  train = sample(1:nrow(VietNamI), 0.8 * nrow(VietNamI))
  test = VietNamI[-train,]
  lnhhexp.lm = lm(lnhhexp ~ age + sex + married + educ +
                    illness + injury + illdays + actdays +
                    insurance + commune, data = VietNamI,
                    subset = train)
  yhat = predict(lnhhexp.lm, newdata = test)
  MSE[i] = mean((yhat - test$lnhhexp) ^ 2)
}
print(MSE)
mean(MSE)
```
```

```
 [1] 0.3362335 0.3362335 0.3362335 0.3362335 0.3362335 0.3362335
 [7] 0.3362335 0.3362335 0.3362335 0.3362335
[1] 0.3362335
```

Utilizing our Linear Regression model, we obtained an average Mean Squared Error (MSE) of 33% across 10 iterations of training and testing using ***lnhhexp*** as our response variable and ***age***, ***sex***, ***married***, ***educ***, ***illness***, ***injury***, ***illdays***, ***actdays***, ***insurance***, and ***commune*** as our predictors. Because MSE measures the average squared difference between the estimated values that our model predicts and the actual results given by the test sample, meaning the best MSE numbers are closest to 0, our model ***lnhhexp.lm*** is fairly accurate.

Bagging (Justin George)

Bagging is a method of aggregating Bootstrap samples to obtain an average prediction value. The bootstrap method is resampling from the original data in order to create many replicate datasets, in order to create proper inferences about specific values related to the data. Bagging is a method of reducing the variance of the typical single decision tree, which often has high variance. In technical terms, we create B individual regression trees from B different bootstrapped training datasets , and we train the method on a specific set from here to get a prediction value. This is repeated and then averaged for a final value. These regression trees are unpruned but large, hence the high variance. We decided to use this model in order to obtain all the benefits of decision trees,
```

importance of variables is no longer clear with so many trees, but the solution to this is using the Gini index to calculate importance.

Model Formula: lnhhexp (predictor) ~ age + sex + married + educ + illness + injury + illdays + actdays + insurance + commune.

While fitting the model, I had to make sure to cast the sex variable to a numeric form, as it was originally categorical in the form of “male” and “female". Additionally, bagging means that every predictor is considered along every split, which increases the amount of time it took for the code to run. The alternative is using a random forest method, which just decreases the amount of predictors per split. Bagging is technically a random forest method.

In creating the bagging model, the number of trees used was a default of 500. The number of predictors used were 11, which included all predictors.



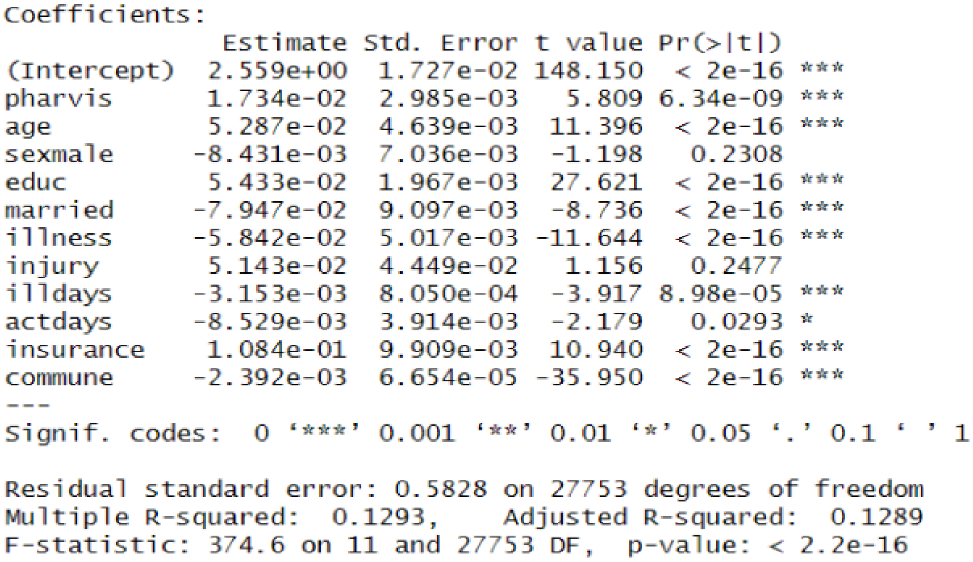
```
Call:
 randomForest(formula = lnhhexp ~ ., data = VietNamI, mtry = 11,      importance = TRUE, keep.forest = TRUE, subset = sample)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 11

          Mean of squared residuals: 0.1029874
                    % Var explained: 73.74
```

The MSE value displayed here is 0.1030. This value will be explained in the results section, additionally with the explanation of how it is obtained.

Linear Regression Results (Riches Dang)

Our response variable is quantitative, the linear regression model is one of, if not the most straight and direct approach to analyzing this data. This method offers simplicity in handling diverse datasets and is straightforward to interpret. Nonetheless, its drawbacks include assuming a linear eliminate outliers as needed, and employ cross-validation to improve predictive accuracy.

Since this is a quantitative data set, we will be making a prediction, by using the test error / training errors. We have the respective coefficients represented below in the summary of our model.

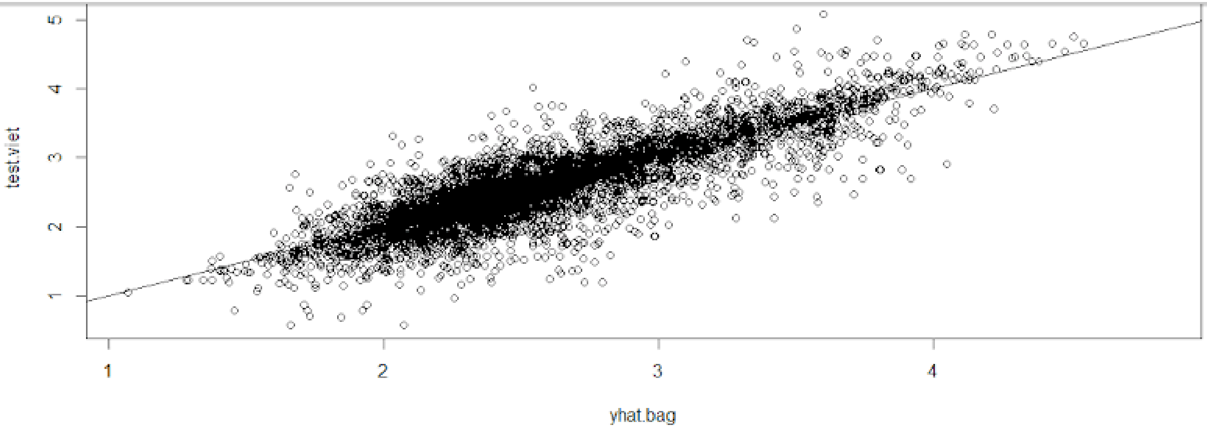


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.559e+00  1.727e-02 148.150  < 2e-16 ***
pharvis      1.734e-02  2.985e-03   5.809 6.34e-09 ***
age          5.287e-02  4.639e-03  11.396  < 2e-16 ***
sexmale     -8.431e-03  7.036e-03  -1.198   0.2308    
educ         5.433e-02  1.967e-03  27.621  < 2e-16 ***
married     -7.947e-02  9.097e-03  -8.736  < 2e-16 ***
illness     -5.842e-02  5.017e-03 -11.644  < 2e-16 ***
injury       5.143e-02  4.449e-02   1.156   0.2477    
illdays     -3.153e-03  8.050e-04  -3.917 8.98e-05 ***
actdays     -8.529e-03  3.914e-03  -2.179   0.0293 *  
insurance    1.084e-01  9.909e-03  10.940  < 2e-16 ***
commune     -2.392e-03  6.654e-05 -35.950  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5828 on 27753 degrees of freedom
Multiple R-squared:  0.1293,	Adjusted R-squared:  0.1289 
F-statistic: 374.6 on 11 and 27753 DF,  p-value: < 2.2e-16
```

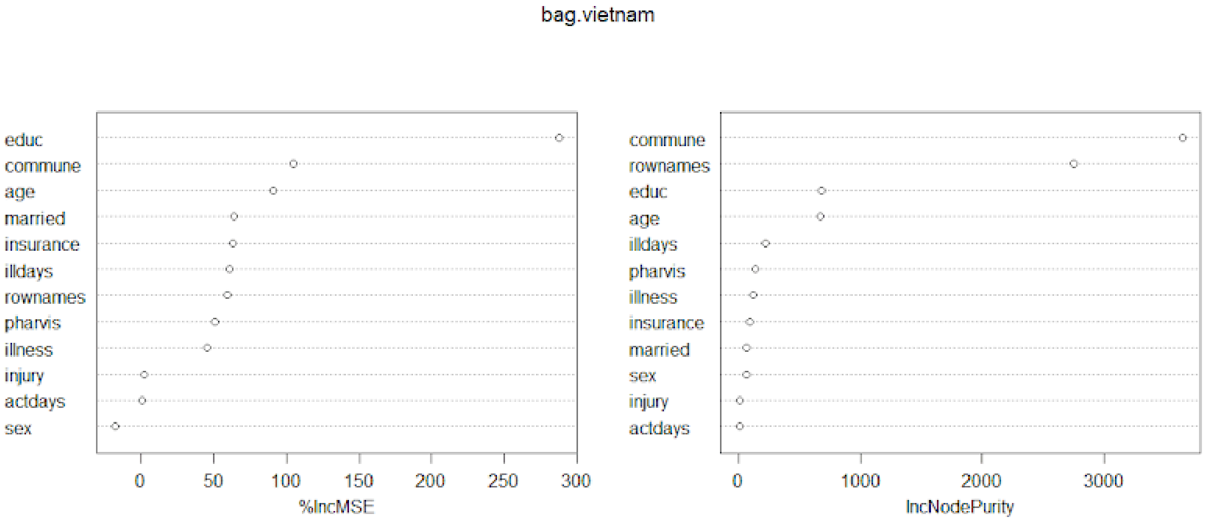
With an F-statistic p-value of less than 2.2 x 10 ^ 16, we can reject the null hypothesis that B1 = B2 = B3 = B4 = B5 = B6 = 0, with these “B”s correlating to the linear regression model equation shown earlier in the report. This shows that at the very least one predictor will have a significant relationship with our response variable. A multiple R^2 value of 0.1293 tells us that 12.93% of the variance is explained by the model, which means this is a horrible fit.

Bagging Results (Justin George)

Through the bagging process, an MSE value of 0.1030 was obtained. To interpret this, we must understand that the response variable is under the operation of a natural logarithm. This means we must undo the operation through the exponential operator, e^X. Therefore, √(e^(0.1030)) = 1.007. This is likely a value scaled down by a thousandth. Undoing this gives us a value of 1007. Therefore we can say that the average test prediction value is within $1007 of the true cost of medical expenditures per family. We can calculate this MSE ourselves, predicting the value from the test set and subtracting it by the true value, and then taking the average of those squared values. This resulted in a value of 0.0979, extremely close to the given MSE.



Additionally, we want to find the importance of each variable as it is not clear with bagging applied. This is done through the Variable Importance Measure plot, shown below. This displays the importance of the education and commune variables. Rownames can be excluded as it is irrelevant.



Conclusion (Riches Dang)

The goal was to understand the economic implications of healthcare expenditure, considering factors such as geographic location, patient demographics, and specific service components. The Linear Regression model, implemented by Chris Do and John Pham, aimed to predict the total amount of money spent on medical expenses based on various predictors. The model considered variables like *pharvis, age, educ, married, illness, illdays, actdays, insurance, and commune*. Through a series of model iterations, the researchers identified the most suitable model (model 2) based on AIC and BIC values. The resulting equation provided insights into the positive and negative effects of model, applied to predictors like age, sex, married, educ, illness, injury, illdays, actdays, insurance, and commune, yielded an MSE value of 0.1030. The interpretation of the MSE indicated that the average test prediction value was within $1007 of the true cost of medical expenditures per family. Riches Dang discussed the Linear Regression results, highlighting the simplicity of the method and its drawbacks, such as assumptions of linearity and susceptibility to outliers. The study emphasized the need for scaling predictor values, outlier elimination, and cross-validation to enhance predictive accuracy. To summarize, the research provided valuable insights into the economic aspects of healthcare in Vietnam, utilizing both Linear Regression and Bagging methods to analyze predictors and their impact on medical expenditures. The findings contribute to a better understanding of the factors influencing healthcare costs for Vietnamese citizens, with implications for policy and decision-making in the healthcare sector.

Bibliography

Cameron, A.C. and P.K. Trivedi (2005) Microeconometrics : methods and applications, Cambridge, pp.848–853.

James, G, Witten, D, Hastie, T, and Tibshirani, R 2013 An Introduction to Statistical Learning. Springer New York. DOI: https://doi.org/10.1007/978-1-4614-7138-7