Contents

*D*ata Analysis Report Project on the Wage Dataset

Tichafa Andrew Rinomhota

Undergraduate in Computer Science, MSU

Executive Summary

The Wage dataset, which is from the ISLR library in R, has an insightful overview of income survey data for males residing in the central Atlantic region of the United States. It acts as a valuable resource for proving the various social and economic factors which affect wages and employment outcomes. This dataset helps us specifically understand the relationships between wages and predictors like age, education level, and job classification, making it a wonderful dataset to show statistical and regression analysis. This dataset consists of 3000 observations and 11 variables, each showing the different attributes on the survey respondents. The variables in this Dataset are Wage, Year, Age, Education, Job Class, Maritl, Race, Health, Healthins, and Logwage. Wage indicates the annual wage in US dollars.Year shows the year the information was recorded.Age shows the age of the worker. Education has levels 1. HS Grad 2. HS Grad 3. Some College 4. College Grad and 5. Advanced Degree indicating education level. Jobclass has levels 1. Industrial and 2. Information indicating type of job. Maritl reflects marital status, including categories like "Never Married", "Married", and "Widowed".Race has categories like "White", "Black", "Asian", and "Other". Health and healthins: Variables denote the perceived health status and health insurance coverage, respectively. The variables in this dataset are a lot and offer categorical and continuous data types, proving there to be multiple analytical approaches.

Wage

Filter

	X	year	age	sex	maritl	race	education	region	jobclass	health
1	231655	2006	18	1. Male	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good
2	86582	2004	24	1. Male	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Gooc
3	161300	2003	45	1. Male	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good
4	155159	2003	43	1. Male	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Gooc
5	11443	2005	50	1. Male	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good
6	376662	2008	54	1. Male	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Gooc
7	450601	2009	44	1. Male	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Gooc
8	377954	2008	30	1. Male	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good
9	228963	2006	41	1. Male	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Gooc
10	81404	2004	52	1. Male	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Gooc
11	302778	2007	45	1. Male	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good
12	305706	2007	34	1. Male	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Gooc
13	8690	2005	35	1. Male	1. Never Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Gooc
14	153561	2003	39	1. Male	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Gooc
15	449654	2009	54	1. Male	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Gooc
16	447660	2009	51	1. Male	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Gooc
17	160191	2003	37	1. Male	1. Never Married	3. Asian	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Gooc
18	230312	2006	50	1. Male	2. Married	1. White	5. Advanced Degree	2. Middle Atlantic	2. Information	2. >=Very Gooc

Data Collection

Of our 3000 entries, we have 11 variables which is a lot of information so we need to provide a summary of our data. From the data, we can see that the minimum age is 18 years, the median age is 42 years, and the maximum age is 80 years. The marital status varaible shows that the most of the individuals are either Never Married (648) or Married (2,074), with a lot less Widowed (19), Divorced (204), and Separated (55). The education variable shows the largest group being High School Graduates (971), followed by individuals with Some College (650), College Graduates (685), and those with Advanced Degrees (426). The racial variable shows that the majority are White (2,480), with smaller proportions of Black (293), Asian (190), and Other (37) individuals. The regional distribution shows that a

lot of individuals are from the Middle Atlantic region (3,000), with no or very few from the other regions listed.

```
     year           age                     maritl
Min.   :2003   Min.   :18.00   1. Never Married: 648
1st Qu.:2004   1st Qu.:33.75   2. Married      :2074
Median :2006   Median :42.00   3. Widowed      :  19
Mean   :2006   Mean   :42.41   4. Divorced     : 204
3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
Max.   :2009   Max.   :80.00

      race                   education
1. White:2480   1. < HS Grad      :268
2. Black: 293   2. HS Grad        :971
3. Asian: 190   3. Some College   :650
4. Other:  37   4. College Grad   :685
                5. Advanced Degree:426


                  region               jobclass
2. Middle Atlantic   :3000   1. Industrial :1544
1. New England       :   0   2. Information:1456
3. East North Central:   0
4. West North Central:   0
5. South Atlantic    :   0
6. East South Central:   0
(Other)              :   0
           health      health_ins      logwage
1. <=Good     : 858   1. Yes:2083   Min.   :3.000
2. >=Very Good:2142   2. No : 917   1st Qu.:4.447
                                    Median :4.653
                                    Mean   :4.654
                                    3rd Qu.:4.857
                                    Max.   :5.763

     wage
Min.   : 20.09
1st Qu.: 85.38
Median :104.92
Mean   :111.70
3rd Qu.:128.68
Max.   :318.34
```

Linear Regression Analysis

To perform linear regression the hypotheses employed would be as follows:
1.) There is no linear significant relationship between "wage" and "age"

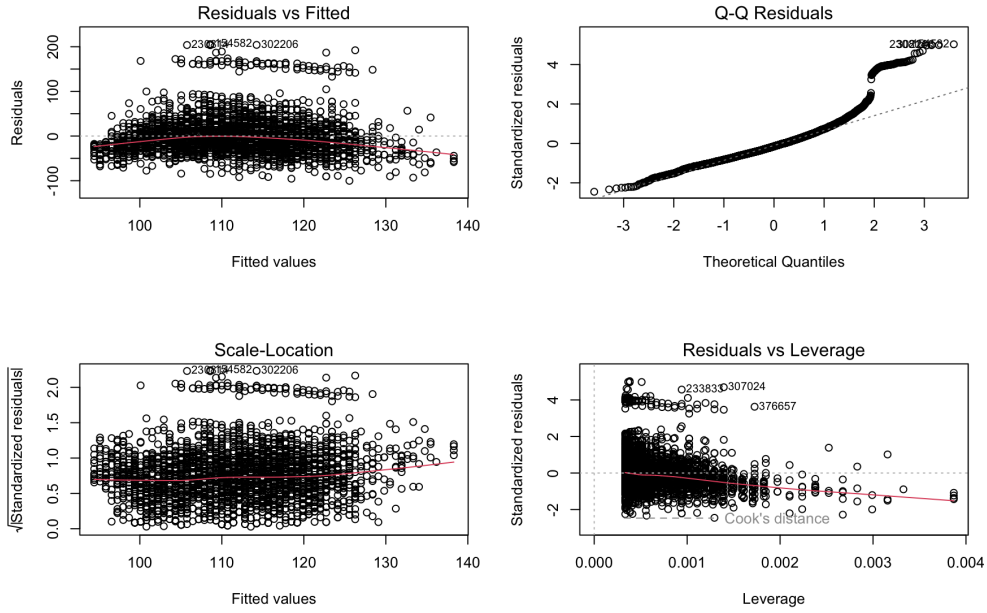
$$H_a : \beta_5 = 0$$

2.) There is a linear significant relationship between "wage" and "age"

$$H_0 : \beta_5 \neq 0$$

The linear regression model reveals a statistically significant relationship between age and wage, as shown by the very small p-value for the age predictor (¡2e-16). The positive coefficient for age (0.70728) suggests that wage increases with age. For every additional year of age, the expected wage rises by approximately 0.707 units. But it looks like, the relationship's strength is relatively weak, as shown by the low R-squared value (0.03827), meaning that only about 3.83 percent of the variance in wages is explained by age. This low value indicates that other varaibles not included in this model likely play a more significant role in determining wages. The diagnostic plots highlight potential concerns with the model fit. The residuals vs. fitted values plot suggests some non-linearity, as the red line deviates from zero across the range of fitted values. The Q-Q plot shows departures from normality in the tails, which might suggest that the residuals are not perfectly normally distributed. The scale-location plot also shows heteroscedasticity. The residuals vs. leverage plot identifies some influential points, though they seem to have a minimal overall effect on the model. The model provides a simple framework to examine the relationship between age and wage, its assumptions suggest it is not sufficient for fully capturing the complexities of wage determination.

```
Call:
lm(formula = wage ~ age, data = Wage)

Residuals:
     Min       1Q   Median       3Q
-100.265  -25.115   -6.063   16.601
     Max
 205.748

Coefficients:
            Estimate Std. Error t value
(Intercept) 81.70474    2.84624   28.71
age          0.70728    0.06475   10.92
            Pr(>|t|)
(Intercept)   <2e-16 ***
age           <2e-16 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 40.93 on 2998 degrees of freedom
Multiple R-squared:  0.03827,   Adjusted R-squared:  0.03795
F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```

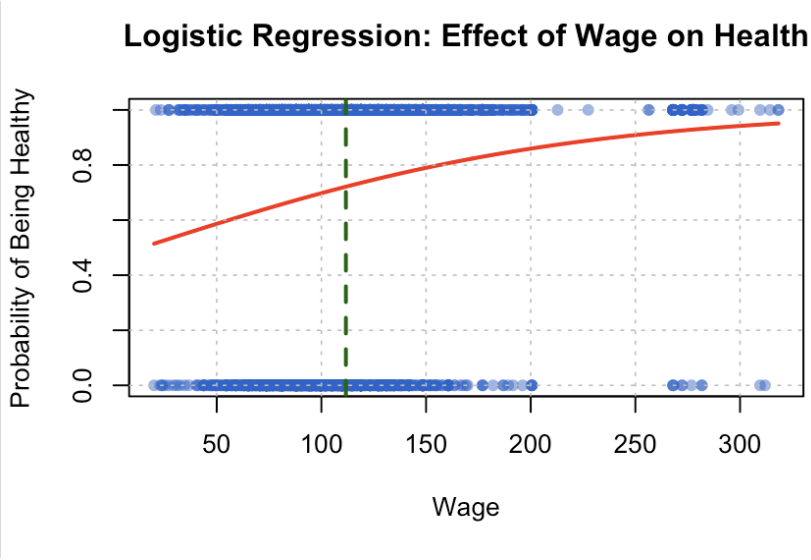


Linear Regression Analysis

The logistic regression model identifies the predictors of health based on the variables wage, health insurance status (healthins), and marital status (maritl). Among the predictors, wage has the strongest and most significant positive effect on health (Estimate = 0.009537, p ¡ 0.001). This shows that as wage increases, the probability of being healthy also increases, which points to idea that higher wages often lead to better access to healthcare, nutritious food, and healthier living conditions. Marital status also has significant effects, with individuals who are divorced having a lower probability of being healthy compared to those who are single (Estimate = -0.577953, p ¡ 0.001). On the other hand, the other marital categories, such as widowed or separated, do not show statistically significant relationships

with health. The variable healthins, which indicates whether someone has health insurance or not, is not significant (p = 0.156). This might suggest that while health insurance is important for accessing healthcare, other factors such as income may have a stronger direct impact on health outcomes. The graph below shows the relationship between wage and the probability of being healthy, as derived from a logistic regression model. The scatter points represent individual observations, with the binary health response (0 or 1) plotted against the corresponding wage values. The red curve represents the logistic regression line, showing the predicted probability of being healthy as a function of wage. The increasing trend of the regression line suggests that higher wages are associated with a greater probabbility of being healthy. The dashed green vertical line marks the mean wage, providing a reference point to the wage distribution. The plot effectively visualizes how wage, is the strongest predictor.

```
Call:
glm(formula = health ~ wage + health_ins + maritl, family = binomial,
    data = Wage)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.102035   0.159939   0.638 0.523500
wage                0.009537   0.001309   7.283 3.25e-13 ***
health_ins2. No    -0.131081   0.092392  -1.419 0.155970
maritl2. Married   -0.181883   0.104755  -1.736 0.082516 .
maritl3. Widowed   -0.470053   0.487366  -0.964 0.334807
maritl4. Divorced  -0.577953   0.170772  -3.384 0.000713 ***
maritl5. Separated -0.276461   0.303273  -0.912 0.361984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3591.2  on 2999  degrees of freedom
Residual deviance: 3499.6  on 2993  degrees of freedom
AIC: 3513.6

Number of Fisher Scoring iterations: 4

>
```



Support Vector Machine

I decided to fit a Support Vector Machine (SVM) classifier using a linear kernel to the Wage dataset, with the aim of predicting the race variable. The ability to predict race in this dataset can provide important insights into demographic patterns and disparities in wages, job classifications, and access to health insurance. Looking at such patterns could help lawmakers, researchers, and employers address systemic inequities or come up with solutions for specific groups. But it is important not to interpret these predictions irresponsibly, for we should avoid discrimination or stereotypes. The race variable is a factor with four levels: "White," "Black," "Asian," and "Other." The SVM was trained with a cost parameter of 0.01, which is used to penalize misclassifications. The model has been evaluated using a train-test split, that I split equally. The model achieved a training accuracy of 82.53 percent, indicating that it correctly predicted the race of approximately 83 percent of the training data. On the test data, the model performed slightly better, with a test accuracy of 82.80 percent. But my training and testing data were low because the model failed to classify any

observations into the "Black," "Asian," or "Other" categories. The kappa statistic, which measures 0, which shows that the model is not performing better than random guessing in terms of agreement with the true values for "Black," "Asian," and "Other. For all non-"White" categories, the model has low sensitivity and high specificity, showing that its not predicting these classes at all. I came to the realization that the model is significantly biased towards predicting the majority class White and this shows class imbalance in the dataset. A lot of the observations belong to the white category, which leads the model to over-predict this class while ignoring the minority classes.While the model performs well for the White class, it fails to identify any instances of "Black," "Asian," or "Other." making it inaccurate.To improve the model's performance, especially for the minority classes, I should consider employing techniques to handle class imbalance. I could over-sample the minority classes or under-sample the majority class to create a more balanced dataset. Or adjust the cost parameter to penalize misclassifications in the minority classes more heavily.I could also perform a grid search to tune the cost and gamma parameters to improve the model's performance and guess for the minorities too.

.

```
Call:
svm(formula = race ~ ., data = train_data,
    kernel = "linear", cost = 0.01)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.01

Number of Support Vectors:  375

 ( 229 9 63 74 )


Number of Classes:  4

Levels:
 1. White 2. Black 3. Asian 4. Other
```

```
Confusion Matrix and Statistics

          Reference
Prediction 1. White 2. Black 3. Asian 4. Other
  1. White     1242      154       85       19
  2. Black        0        0        0        0
  3. Asian        0        0        0        0
  4. Other        0        0        0        0

Overall Statistics

               Accuracy : 0.828
                 95% CI : (0.8079, 0.8468)
    No Information Rate : 0.828
    P-Value [Acc > NIR] : 0.5166

                  Kappa : 0
```

Conclusion

Summary of Key Findings This analysis of the Wage dataset revealed several important insights. Firstly, a linear regression showed a statistically significant but weak relationship between age and wages, showing that while wages tend to increase with age, other factors not captured in this model play a larger role in determining wage outcomes. A logistic regression model highlighted that wage is the strongest predictor of health, with

higher wages associated with a bigger chance of being healthy, while health insurance status had no significant effect. The Support Vector Machine classifier aimed at predicting race which showed a high accuracy for the White category but failed to predict the minority categories proving a significant class imbalance in the dataset.

Implications of the Results The findings suggest that economic factors, particularly wages, are strongly tied to health outcomes, emphasizing how important wage equal access to healthcare are. The weak relationship between age and wages indicates that other social economic factors may be more influential in wage determination. The SVM's inability to predict minority racial categories points to potential biases in the dataset or how this region is heavily populated by one race.But it could also mean disparities in job classification and wage distribution.

1 Appendix

- R code

```
> library(ISLR)
> data(Wage)
> model <- lm(wage ~ age, data = Wage)
> summary(model)

> plot(Wage$age, Wage$wage, main = "Wage vs Age", xlab = "Age", ylab = "Wa
> abline(model, col = "red", lwd = 2)
> par(mfrow = c(2, 2))  # Arrange plots in a 2x2 grid
> plot
> logit_model <- glm(health ~ wage + health_ins + maritl,
+                    data = Wage, family = binomial)
> summary(logit_model)

> library(e1071)
> library(caret)
> data("Wage", package = "ISLR")
> Wage$race <- as.factor(Wage$race)
> str(Wage)
> Wage_scaled <- Wage
> numeric_columns <- which(sapply(Wage, is.numeric))  # Find the numeric c
> Wage_scaled[, numeric_columns] <- scale(Wage[, numeric_columns])
> set.seed(42)
> train_index <- sample(1:nrow(Wage_scaled), nrow(Wage_scaled) / 2)
> train_data <- Wage_scaled[train_index, ]
> test_data <- Wage_scaled[-train_index, ]
> svm_model <- svm(race ~ ., data = train_data, kernel = "linear", cost =
> summary(svm_model)
> train_predictions <- predict(svm_model, newdata = train_data)
> train_accuracy <- mean(train_predictions == train_data$race)
> print(paste("Training Accuracy:", round(train_accuracy, 4)))
> test_predictions <- predict(svm_model, newdata = test_data)
> test_accuracy <- mean(test_predictions == test_data$race)
> print(paste("Test Accuracy:", round(test_accuracy, 4)))
```

```
> confusion_matrix <- confusionMatrix(test_predictions, test_data$race)
> print(confusion_matrix)
```

References: rds.had.co.nz, StackOverflow,