METCS 544 Final Project

Introduction

Obesity has emerged as a critical public health issue, and understanding the factors that contribute to varying levels of obesity can guide both preventive strategies and clinical interventions. The Estimation of Obesity Levels dataset from the UC Irvine Machine Learning Repository includes measurements of weight, height, age and a detailed obesity classification variable called NObeyesdad. This classification is based on each individual's BMI (kg/m^2), calculated from weight and height, then categorized using established thresholds: BMI <18.5 as Insufficient Weight, 18.5–24.9 as Normal Weight, 25.0–29.9 as Overweight, and ≥30 as Obesity. To be more accurate, 30.0-34.9 will be defined as Obesity I. 35.0-39.9 will be defined as Obesity II and higher than 40 will be defined as Obesity III. And Overweight is separated into two groups. This report examines two questions that illuminate key dimensions of this problem. The first explores whether a simple linear regression can describe the relationship between an individual's height and weight. The second investigates how age varies across different BMI group.

Methods

The primary analytical approaches are simple linear regression (SLR) and one-way analysis of variance (ANOVA). For SLR, after taking a random sample of 200, height (in centimeters) served as the predictor and weight (in kilograms) as the response; First step is to set up hypothesis. In this case, the null hypothesis would be $\beta = 0$, and alternative hypothesis is β not equal to 0. Use t-test to test the significance of the slope. Use both p-value and 95% confidence interval to make decision about our hypothesis. And also make sure that the conditions for inference are satisfied — linearity, independence, normality, equal SD and random sample —ensured validity of regression assumptions before interpretation.

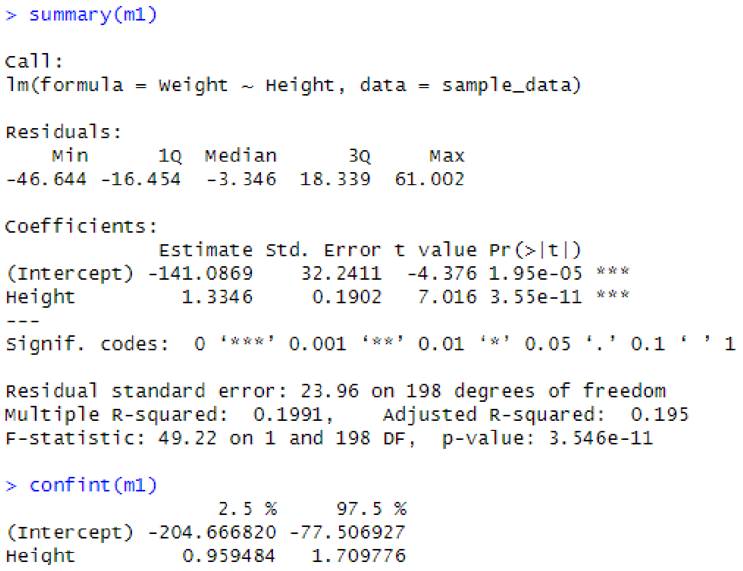
For ANOVA, different types of obesity labels were first grouped into Insufficient Weight, Normal Weight, Overweight, and Obesity categories according to the original group. Mean age is compared between groups, and different BMI group is used as a categorical factor. The F-test of the ANOVA assessed whether mean ages differed across groups, and using Tukey to specific pairwise differences. And we also need to check the assumptions to perform ANOVA, which are independence, normality, equal variance and measurement scale. If these assumptions are not correct, we need to transform the data into a form that meet these requirements.

Results

Simple Linear Regression

First, I transformed the height into centimeters.

Applying simple linear regression, the estimated relationship between height and weight was: Weight = -141.09 + 1.33 × Height. The slope coefficient was highly significant (t = 7.016, p-value = 3.55e-11), with a 95% confidence interval of (0.96, 1.71). The model accounted for 19.9% of the variance in weight ($R^2 = 0.1991$).



```
> summary(m1)

Call:
lm(formula = Weight ~ Height, data = sample_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.644 -16.454  -3.346  18.339  61.002 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -141.0869    32.2411  -4.376 1.95e-05 ***
Height         1.3346     0.1902   7.016 3.55e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.96 on 198 degrees of freedom
Multiple R-squared:  0.1991,    Adjusted R-squared:  0.195 
F-statistic: 49.22 on 1 and 198 DF,  p-value: 3.546e-11

> confint(m1)
                  2.5 %     97.5 %
(Intercept) -204.666820 -77.506927
Height         0.959484   1.709776
```

Figure 1

About the conditions for inference. We have this scatterplot of roughly linear pattern.

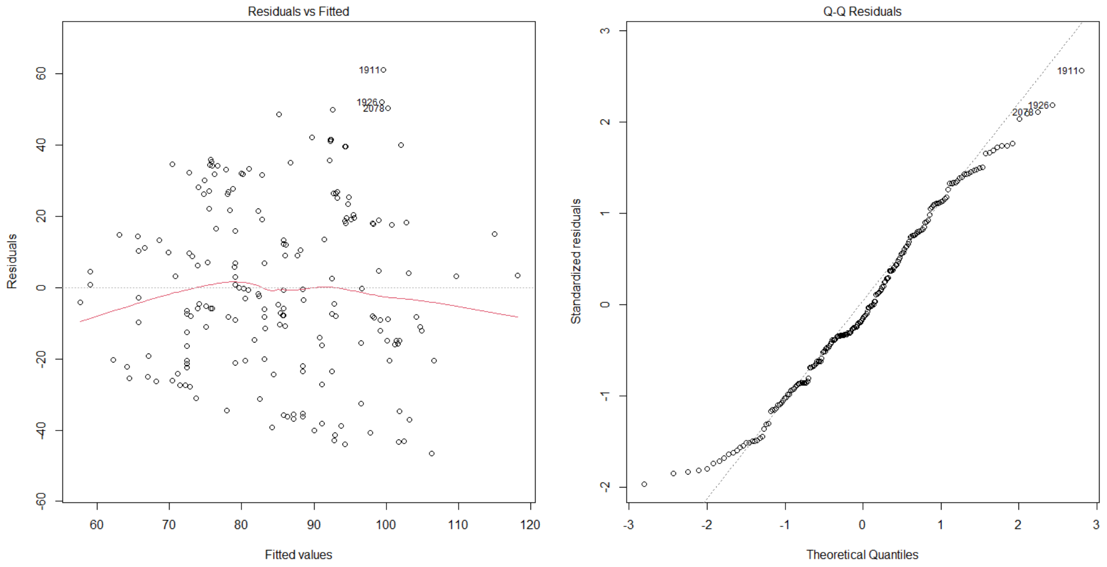


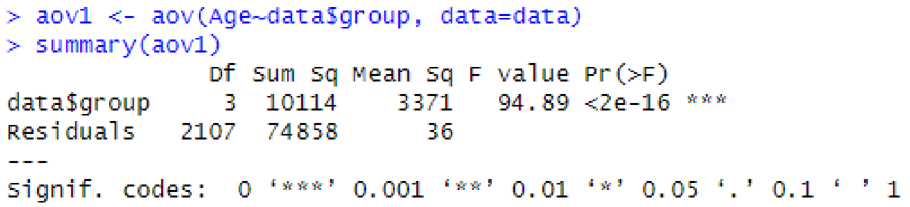
Figure 2

And we have 200 smaller than 10%*2111, so the data represents an independent sample. The Q-Q plot (Figure 2) shows the normality is satisfied. And from the scatter plot, we can tell it shows a random scatter around the residual = 0 line. Finally, since our sample is randomly selected, so we can reject the null hypothesis. We have convincing evidence of a positive linear relationship between weight and height.

Analysis of variance

The sample included four categorical groups: Insufficient Weight, Normal Weight, Overweight, and Obesity. I combined two overweight types into one single group and three obesity types into one single group. We'll have to assume that the independence assumption is satisfied and the age is measured on a ratio scale.

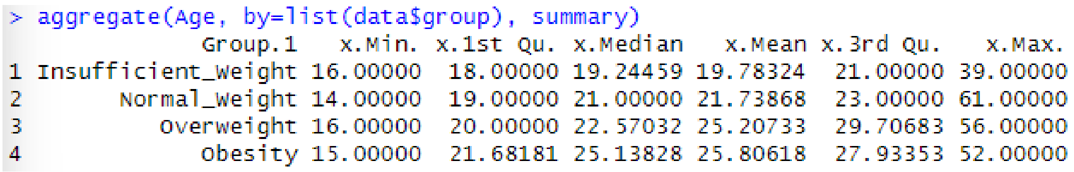
A one-way ANOVA on age produced F statistic of 94.89 (p < 2e-16), indicating there's at least one significant group differences (Figure 3)



```
> aov1 <- aov(Age~data$group, data=data)
> summary(aov1)
                 Df Sum Sq Mean Sq F value Pr(>F)    
data$group        3  10114    3371   94.89 <2e-16 ***
Residuals      2107  74858      36                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3

And using the aggregate function, we have overweight and obesity group's mean are very similar. (Figure 4)



```
> aggregate(Age, by=list(data$group), summary)
              Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
1 Insufficient_Weight 16.00000  18.00000 19.24459 19.78324  21.00000 39.00000
2       Normal_Weight 14.00000  19.00000 21.00000 21.73868  23.00000 61.00000
3          Overweight 16.00000  20.00000 22.57032 25.20733  29.70683 56.00000
4             Obesity 15.00000  21.68181 25.13828 25.80618  27.93353 52.00000
```

Figure 4

But if we take a look at the Q-Q plot and scale location-plot (Figure 5), we can see that the normality and constant variance assumption are not satisfied.

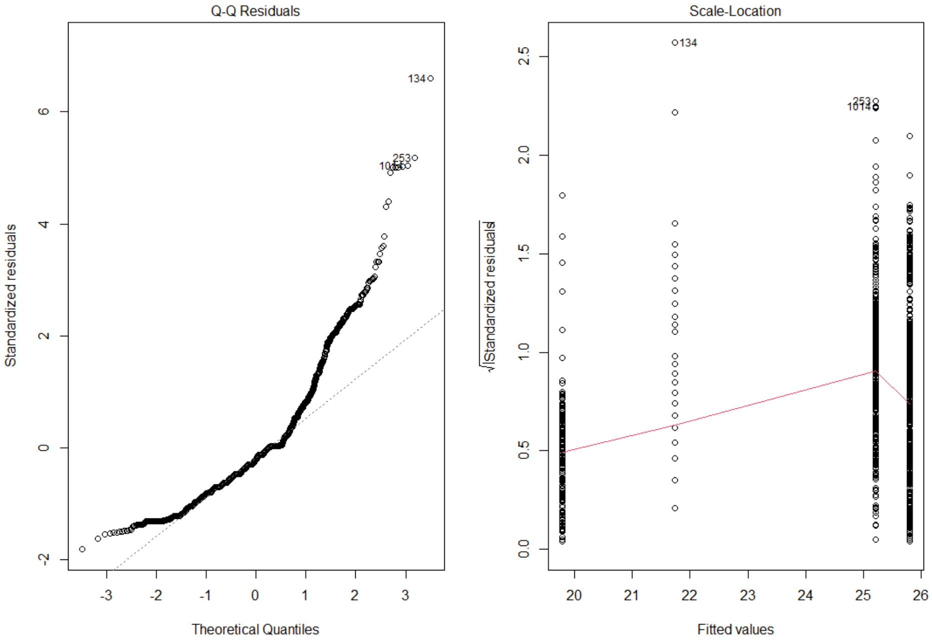


Figure 5

So we used Box-Cox to find the optimal Box-Cox transformation index lamda for Age, and convert the original age into Age_bc and store it back in the dataset. Then we used the transformed age to fit the linear model and took a look at the Q-Q plot and residual plot (Figure 6).

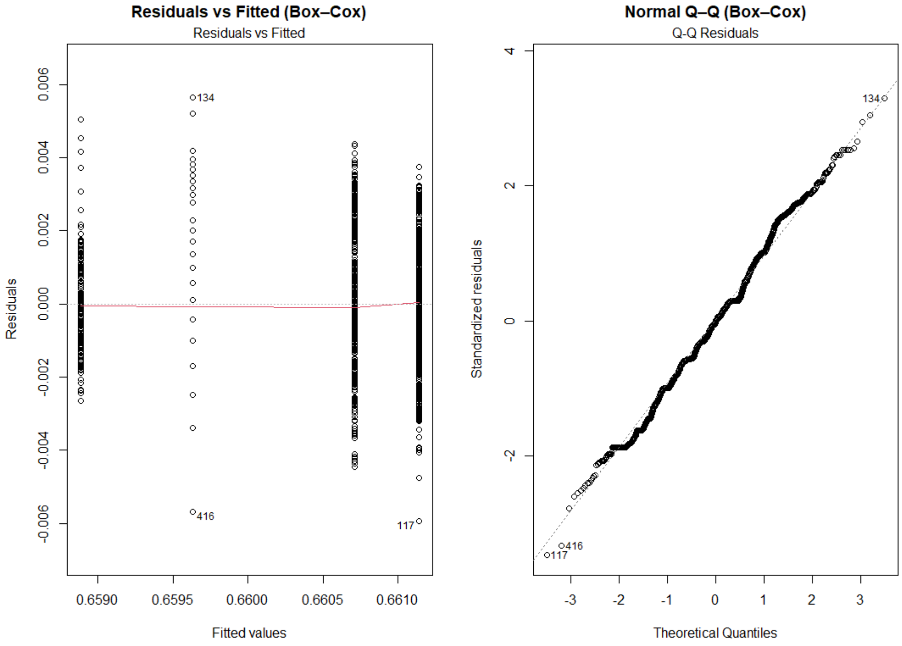
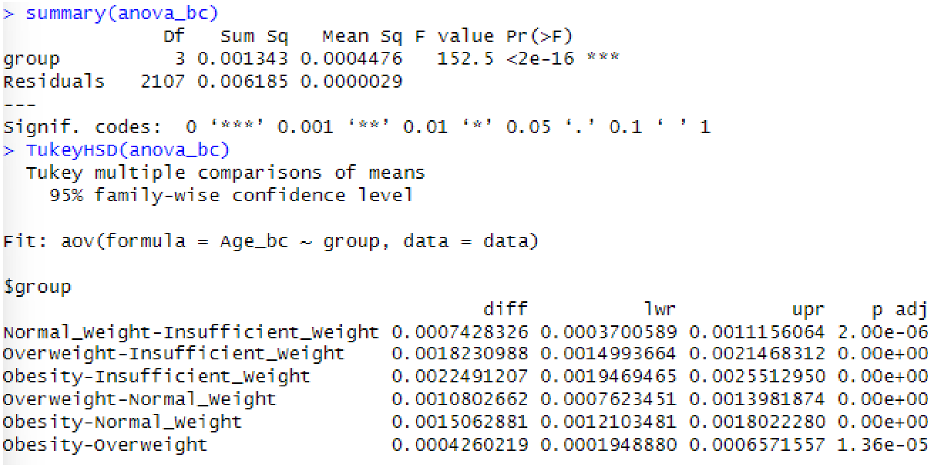


Figure 6

These two graphs generally conform to the diagnostic criteria of normality and equal variance, which indicates that the Box-Cox transformation effectively improves the model assumption. Finally, we performed ANOVA and got the F statistics of 152.5 with p-value smaller than 2e-16, which is really significant, indicating that there's at least two groups whose mean ages after adjusting are different.



```
> summary(anova_bc)
              Df   Sum Sq   Mean Sq F value Pr(>F)    
group          3 0.001343 0.0004476   152.5 <2e-16 ***
Residuals   2107 0.006185 0.0000029                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova_bc)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Age_bc ~ group, data = data)

$group
                                      diff          lwr          upr    p adj
Normal_Weight-Insufficient_Weight 0.0007428326 0.0003700589 0.0011156064 2.00e-06
Overweight-Insufficient_Weight    0.0018230988 0.0014993664 0.0021468312 0.00e+00
Obesity-Insufficient_Weight       0.0022491207 0.0019469465 0.0025512950 0.00e+00
Overweight-Normal_Weight          0.0010802662 0.0007623451 0.0013981874 0.00e+00
Obesity-Normal_Weight             0.0015062881 0.0012103481 0.0018022280 0.00e+00
Obesity-Overweight                0.0004260219 0.0001948880 0.0006571557 1.36e-05
```

Figure 7

To further examine the sources of differences in the mean age (after Box-Cox transformation) among the four BMI groups, we conducted a Tukey HSD test on the ANOVA model Age_bc ~ group. The adjusted P-values among all groups were much less than 0.05, indicating that among the four BMI groups from "underweight" to "normal weight" and then to "overweight" and "obese", there were highly significant pairwise differences in the average age after transformation. Finally, the 95% family-wise plot after transforming will look like this:

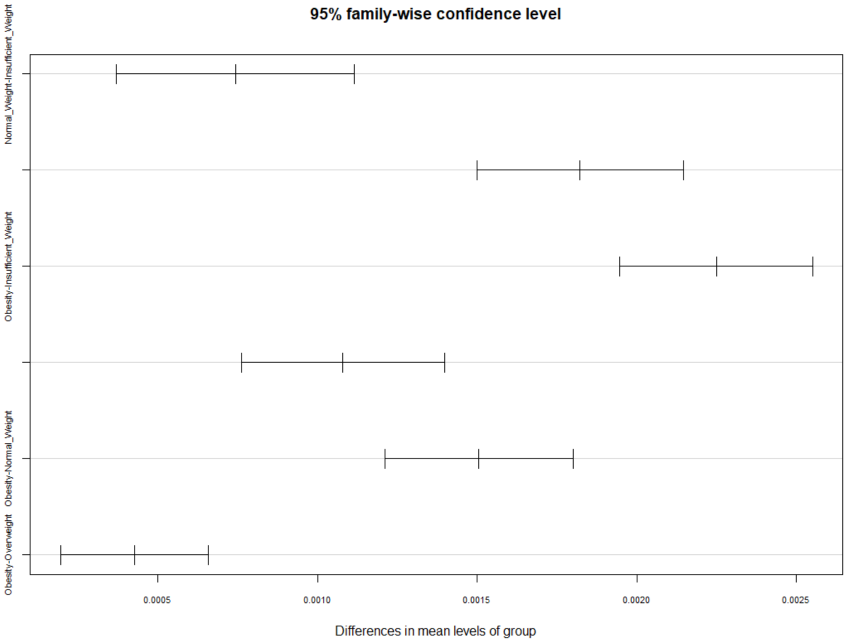


Figure8

Discussion and Conclusion

The regression results demonstrate that each additional centimeter in height corresponds to approximately 1.33 kg in weight, with a narrow 95% confidence interval (0.96, 1.71). Although the association is statistically strong ($p < 3.55e-11$), the R^2 of 0.1991 indicates that height alone explains just over one-fifth of weight variability. In practical terms, this suggests that while taller individuals in this population do weigh more on average, factors such as diet and activity levels likely play roles. In the project, I don't include variables like exercise habits or diet. A multi-linear model may better explain weight variation.

In the age-by-group comparison, Initial ANOVA on untransformed Age violated normality and homogeneity of variance. After applying a Box–Cox transformation, both the Q–Q plot and residuals–fitted plot demonstrated approximate normality and constant variance. We noticed that mean age increases monotonically across BMI categories, suggesting that older individuals are progressively more likely to fall into higher BMI classes. This may reflect cumulative lifestyle and metabolic changes with age. And the public health should emphasize healthy weight maintenance in middle-aged and older adults, while monitoring underweight risks in younger people. The Box–Cox step proved critical to meet ANOVA assumptions; without it, standard F–tests risk inflated Type I error or reduced power under different variance group.

In summary, our simple linear regression confirms a positive, statistically significant but weak relationship between height and weight, highlighting that height explains only a portion of weight variability. One-way ANOVA on Box–Cox–transformed age reveals highly significant differences in mean age across BMI categories: as weight classification moves from underweight to obese, the average age steadily increases. These findings underscore the importance of age-sensitive strategies in obesity prevention and control, and demonstrate the necessity of appropriate data transformations to fulfill statistical model assumptions.

Video

https://youtu.be/vmQbCHSUNko

References

UCI Machine Learning Repository. Estimation of Obesity Levels based on Eating Habits and Physical Condition. https://archive.ics.uci.edu/dataset/544

Mendoza Palechor, F., & De la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico [Data set]. *Data in Brief, 25*, 104344. https://doi.org/10.1016/j.dib.2019.104344