STAT363 HOMEWORK 2

Berfin AKDEMİR/2614410, Zeynep KANAR/2561306

2024-08-11

STAT363 LINEAR MODELS I HOMEWORK 2

INTRODUCTION

This research investigates the impact of biking to work and smoking on heart disease using a dataset of 498 rows simplified for practicality. The dataset includes the following variables:

Percentage of People Biking to Work: This represents the proportion of individuals biking to work daily. This is a controllable variable, as it can be influenced by interventions such as promoting cycling. Dependent Variable:

Percentage of People with Heart Disease: This shows the proportion of individuals with heart disease in each town. This is the outcome variable influenced by various factors including biking and smoking. Uncontrollable Variable:

Percentage of People Smoking: Reflects the proportion of people who smoke. While it is a key factor, it is less controllable on an individual level compared to biking. Smoking rates can be influenced by public health policies but are not easily controlled by individuals.

The key research questions in this project are:

How does the percentage of people biking to work influence the risk of heart disease? What is the effect of the percentage of people smoking on the incidence of heart disease?

Understanding these relationships is important as it helps identify key factors that influence health outcomes. Insights from this analysis can inform individual choices and guide public health strategies.

R libraries used in this study:

```r
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)
library(car)
library(lmtest)
```

ANALYSIS

Let’s start by reading and observing the data

```r
data <- read.csv("/Users/berfinakdemir/Desktop/STAT363/Homework/HW2/heart.data.csv")

head(data)
```

```
##   X    biking   smoking heart.disease
## 1 1 30.801246 10.896608     11.769423
## 2 2 65.129215  2.219563      2.854081
## 3 3  1.959665 17.588331      17.177803
## 4 4 44.800196  2.802559       6.816647
## 5 5 69.428454 15.974505       4.062224
## 6 6 54.403626 29.333176       9.550046
```

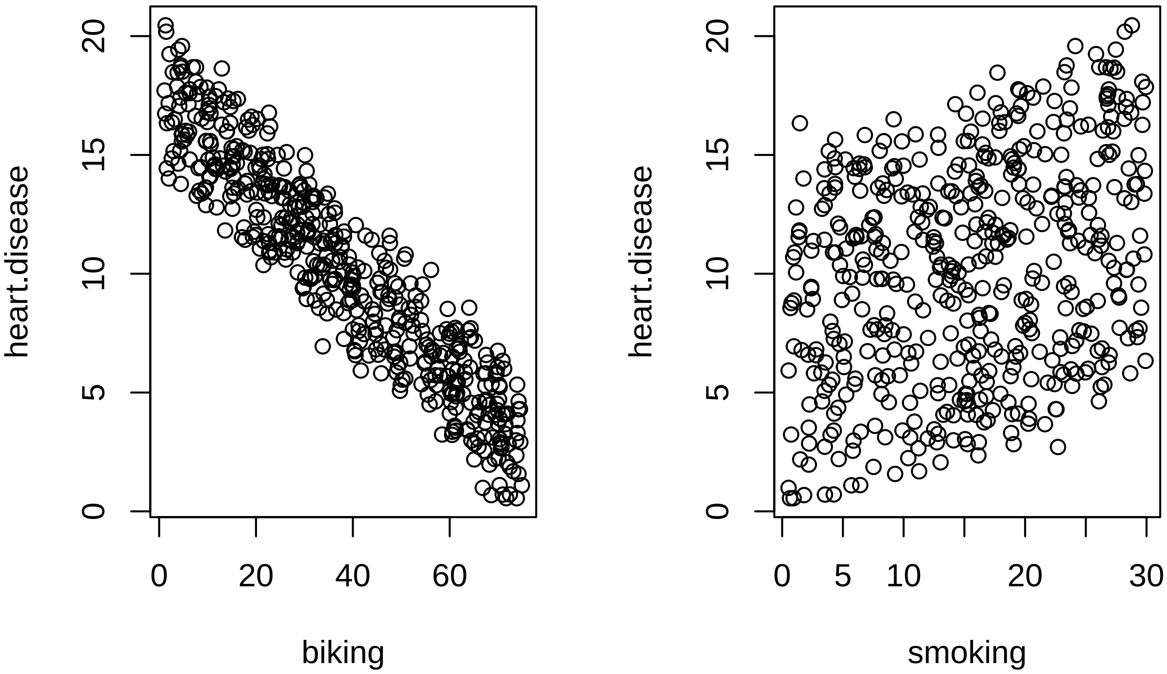
```
str(data)
```

```
## 'data.frame':    498 obs. of  4 variables:
##  $ X            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ biking       : num  30.8 65.13 1.96 44.8 69.43 ...
##  $ smoking      : num  10.9 2.22 17.59 2.8 15.97 ...
##  $ heart.disease: num  11.77 2.85 17.18 6.82 4.06 ...
```

MULTIPLE LINEAR REGRESSION ASSUMPTIONS

Linearity of the data: The relationship between the predictor (x) and the outcome (y) is assumed to be linear. We observed scatter plots.

```
par(mfrow = c(1, 2))
plot(heart.disease ~ biking, data=data)
plot(heart.disease ~ smoking, data=data)
```

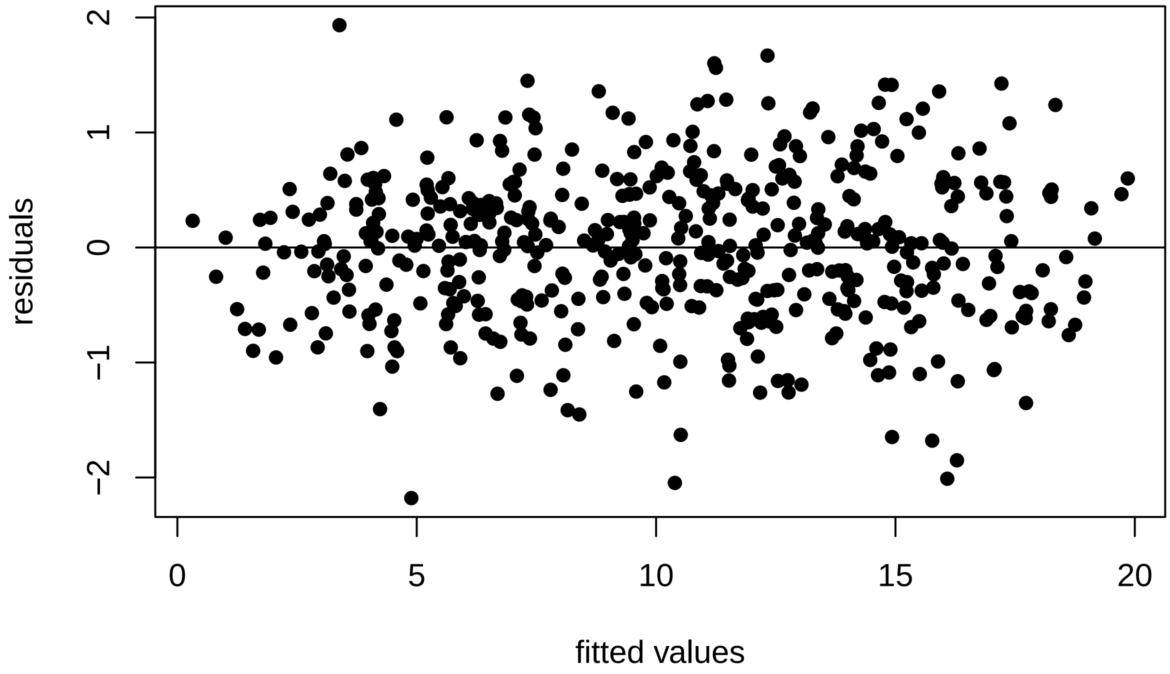


From the plots we observe negative linear relationship with biking and a weak positive linear relationship with smoking. We proceed to check the other assumptions by residuals since solely this plots are not sufficient.

```
model <- lm(heart.disease~ biking+smoking ,data=data)
```

Homogeneity of residuals variance: The residuals are assumed to have a constant variance (homoscedasticity) We observe a plot between residuals and fitted values to see if the variances are equally distributed.

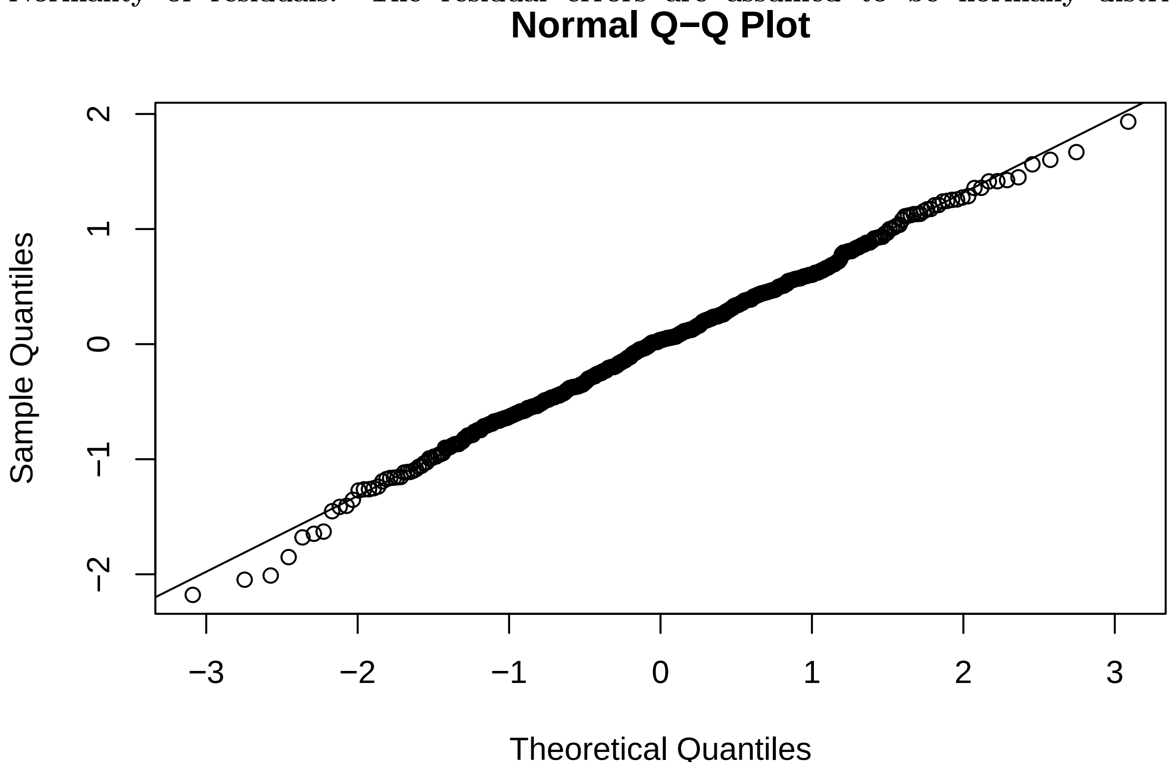
```
yhat=predict(model)  ### obtaining yhat by predict()
resid=model$residuals
plot(yhat,resid,pch=16,xlab="fitted values",ylab="residuals")
abline(h=0)
```



The residuals randomly scattered around the fitted line which implies equal variance between them.

```
resid=model$residuals
qqnorm(resid)
qqline(resid)
```

Normality of residuals: The residual errors are assumed to be normally distributed.



```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.997, p-value = 0.4935
```

The values on the qq-plot lies on the line and Shapiro-Wilk test for residuals has a p-value of 0.4935 which is clearly larger than 0.05. Therefore we can conclude that errors are normallly distributed.

Independence of residuals error terms. (Multicollinearity) Variance Inflation Factor and Durbin - Watson test is used to observe the relationship between residuals error terms.

```
vif(model)
```

```
##   biking  smoking
## 1.000229 1.000229
```

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.9174, p-value = 0.1773
## alternative hypothesis: true autocorrelation is greater than 0
```

VIF is seen to be smaller than 5 and Durbin-Watson test results show a p-value larger than 0.05, we conclude that residual error terms are independent and there is no multicollinearity problem.

HYPOTHESIS & TESTING

ANOVA Examining an ANOVA on the model we see:

H0 = model is not significant Ha = model is significant

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: heart.disease
##            Df Sum Sq Mean Sq F value    Pr(>F)
## biking      1 9090.6  9090.6 21251.7 < 2.2e-16 ***
## smoking     1 1086.0  1086.0  2538.8 < 2.2e-16 ***
## Residuals 495  211.7     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On ANOVA we have contemplated p-values which are both smaller than 0.05 which implies that both of the regressors are significant.

```

summary(model)
```

General Observation of the model

```
##
## Call:
```

```
## lm(formula = heart.disease ~ biking + smoking, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1789 -0.4463  0.0362  0.4422  1.9331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.984658   0.080137  186.99   <2e-16 ***
## biking      -0.200133   0.001366 -146.53   <2e-16 ***
## smoking      0.178334   0.003539   50.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.654 on 495 degrees of freedom
## Multiple R-squared:  0.9796, Adjusted R-squared:  0.9795
## F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

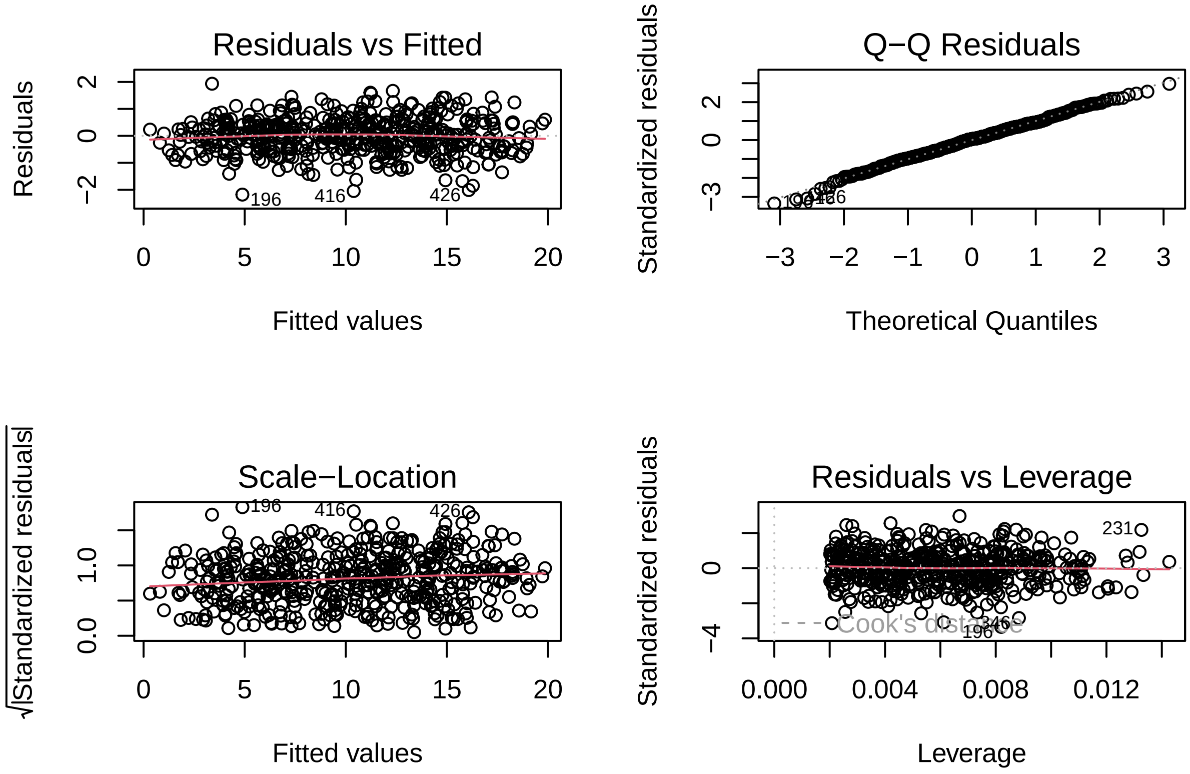
We examine the summary to furtherly see the change of ß1 (biking) and ß2 (smoking) on the model.

The estimate value for the biking shows that there is a negative relationship with the outcome variable since it has a negative sign, with smoking sign of the coefficient implies a positive relationship with the proportion of the hear disease. Both of the variables again are considered significant for the model with clearly lower than error (0.05) p-values.

We also have high coefficient of determination value that points out ~%98 of the distribtion of the sample is explained in our model.

A more profound explanation of these computations are stated in conclusion part.

```
par(mfrow=c(2,2))
plot(model)
```



Residuals vs Fitted

Q−Q Residuals

Scale−Location

Residuals vs Leverage

For further understanding of the model; scale-location plot has standardized residual values bouncing randomly around the fitted line which implies homoscedasticity of the distribution, residual-leverage plot shows no significant outliers.

CONCLUSION

Throughout this study, we wanted to observe the effect of biking and smoking on the proportion of people with heart disease. The data was simplified for practicality before this study, so there was no need to clean it. Firstly, we observed assumptions of linear regression.

Linearity of the data: we would expect a linear relationship between the outcome and regressors, the simplest way to achieve that is creating scatter plots. We thought scatter plots showed that smoking has a slightly increasing effect on heart diseases, and biking has a decreasing effect.

Homogeneity of residuals variance: The residuals are assumed to have constant variance (homoscedasticity): To observe the change of variances clearly, we checked the residuals vs fitted values plot to see how the values act around the 0 line. Since they bounce randomly around the line, it is possible the understand that the assumption is satisfied.

Normality of residuals: The residual errors are assumed to be normally distributed: The values follow the line on a normal QQ plot, therefore they have adopted a normal distribution. To be sure, we have also conducted a Shapiro test which also resulted in a p-value larger than our error.

Independence of residuals error terms (Multicollinearity): To observe the interaction between residual error terms we checked VIF (variance inflation factor) and conducted a Durbin-Watson test. VIF values were both smaller than 5, and the Durbin-Watson test had a p-value of $0.1773 > 0.05$. That concluded our assumptions and we continued with our model.

After making sure that all assumptions were satisfied, we conducted ANOVA and a summary of the model. Both ß1 and ß2 are seen to be significant coefficients through the ANOVA and the summary. Mainly we can say that the higher proportion of biking resulted in less numbers of heart diseases, and smoking has an increasing effect on proportion of heart disease.

FUTURE DIRECTIONS

If we had the opportunity to conduct this study without any limitations on data sources, we would enhance our analysis by incorporating variables such as genetic factors, demographic characteristics (age, gender, ethnic background, income level, marital status, occupation), geographic location (urban vs. rural living), dietary habits (daily intake of fats, proteins, carbohydrates, vitamins, minerals, water, alcohol, salt, and the quality of these nutrients), individual stress levels, sleep quality, psychological state, and environmental factors (air pollution, noise levels, climate).

Additionally, collecting data from a large population equipped with sensors that continuously monitor health conditions could enable real-time prediction of health risks. This would facilitate early intervention for diseases and allow for personalized health strategies tailored to individual needs.