Simple and Multiple Linear Regression Report

Biola Madandola

Simple Linear Regression

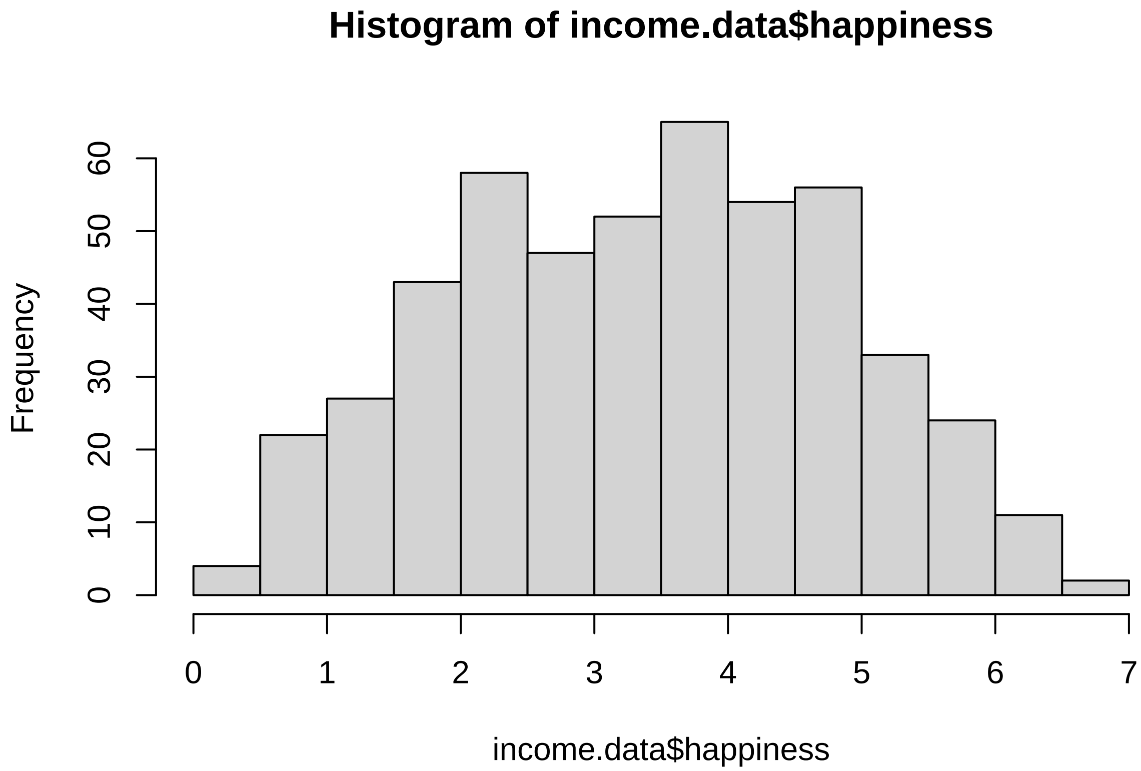
Load and Summarize Data

```r
income.data <- read.csv("income.data.csv")
summary(income.data)
```

```
##        X             income        happiness
##  Min.   :  1.0   Min.   :1.506   Min.   :0.266
##  1st Qu.:125.2   1st Qu.:3.006   1st Qu.:2.266
##  Median :249.5   Median :4.424   Median :3.473
##  Mean   :249.5   Mean   :4.467   Mean   :3.393
##  3rd Qu.:373.8   3rd Qu.:5.992   3rd Qu.:4.503
##  Max.   :498.0   Max.   :7.482   Max.   :6.863
```

Histogram of Happiness

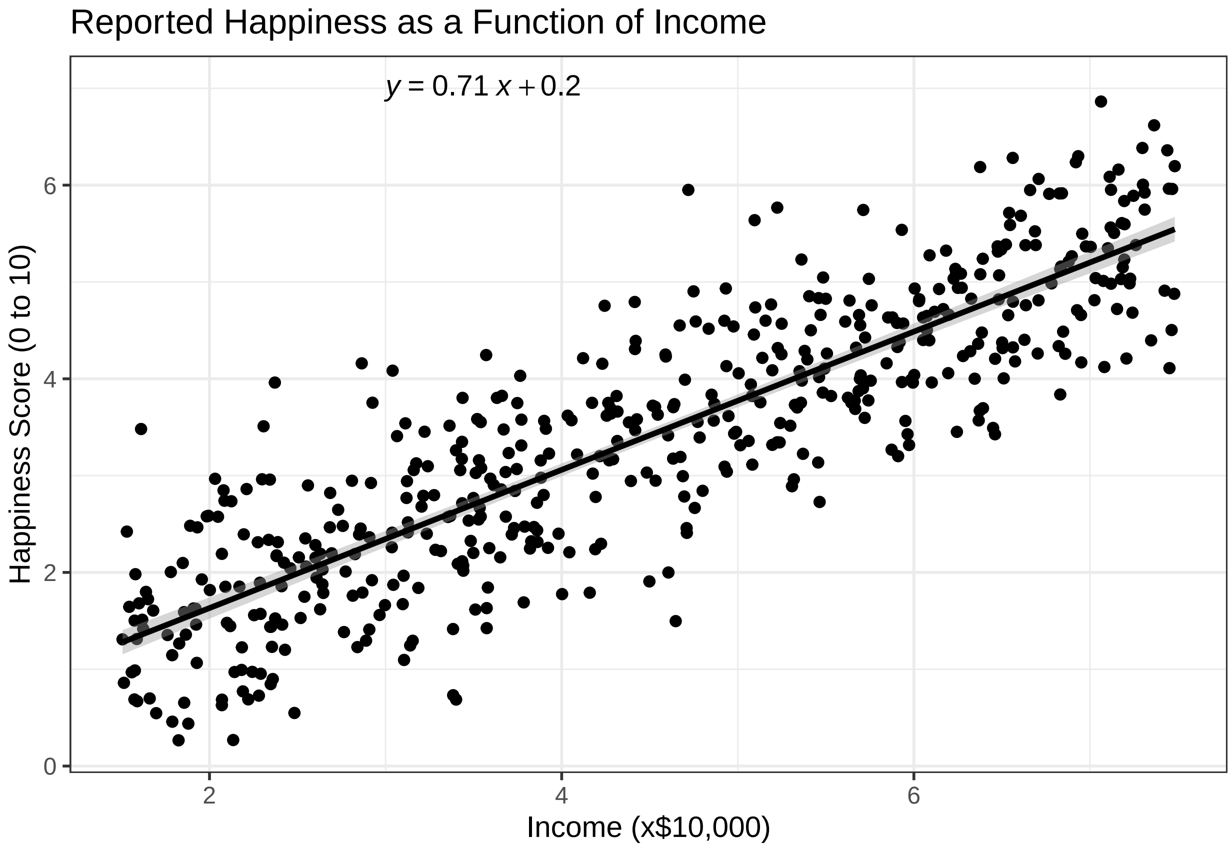
```r
hist(income.data$happiness)
```



Scatter Plot and Regression Line

```
income.graph <- ggplot(income.data, aes(x = income, y = happiness)) +
  geom_point() +
  geom_smooth(method = "lm", color = "black") +
  stat_regline_equation(label.x = 3, label.y = 7) +
  theme_bw() +
  labs(
    title = "Reported Happiness as a Function of Income",
    x = "Income (x$10,000)",
    y = "Happiness Score (0 to 10)"
  )
income.graph
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Regression Summary

```
income.happiness.lm <- lm(happiness ~ income, data = income.data)
summary(income.happiness.lm)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = income.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

Interpretation

The simple linear regression shows a strong positive relationship between income and happiness (R^2 = 0.7493). The slope coefficient of 0.714 suggests that, on average, a $10,000 increase in income is associated with a 0.714 point increase in happiness score.

Multiple Linear Regression

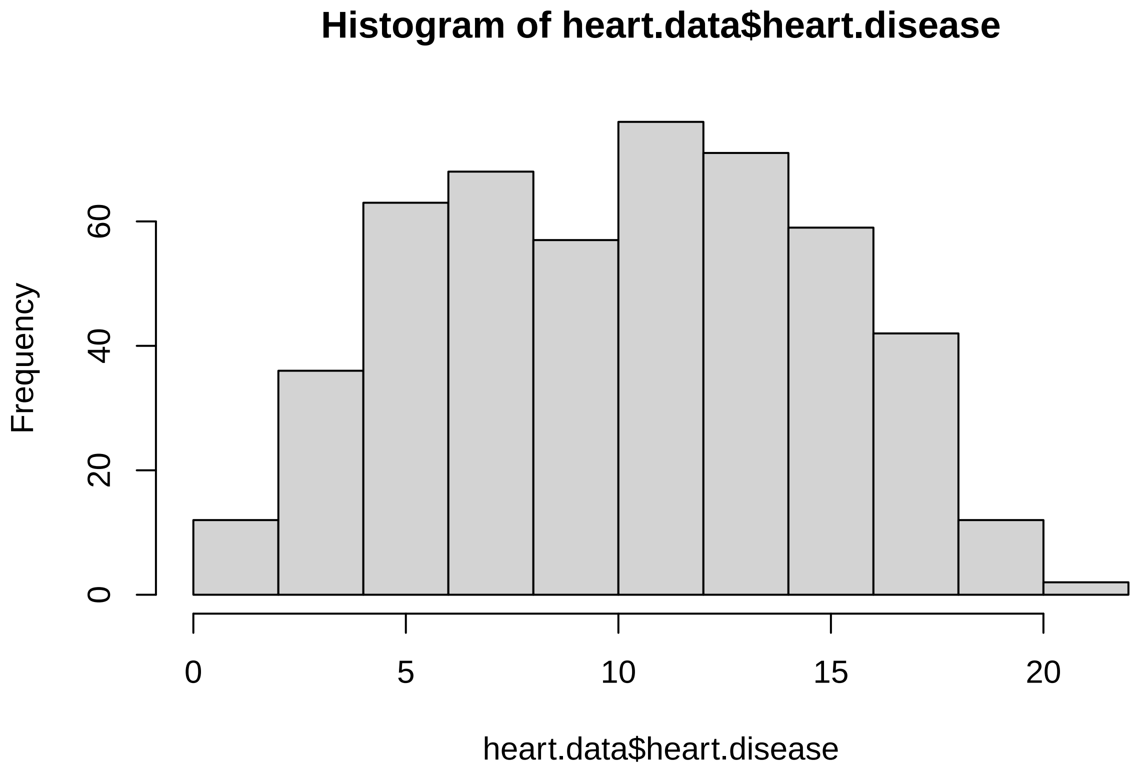
Load and Summarize Data

```
heart.data <- read.csv("heart.data.csv")
summary(heart.data)
```

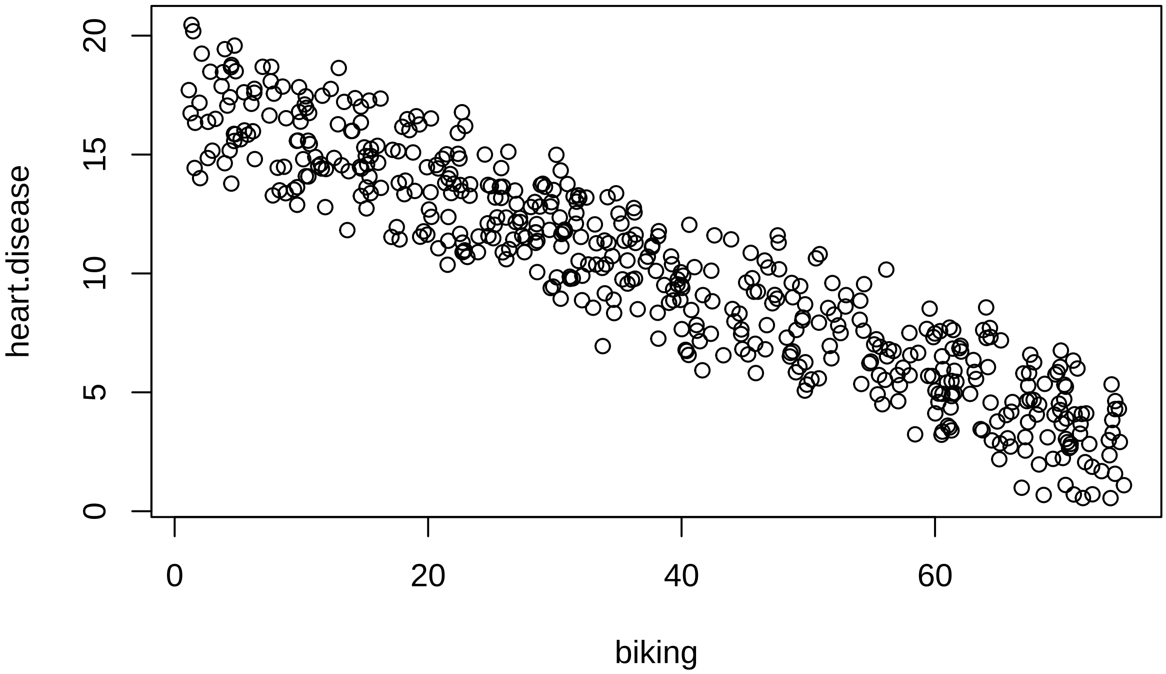
```
##        X             biking          smoking        heart.disease
##  Min.   :  1.0   Min.   : 1.119   Min.   : 0.5259   Min.   : 0.5519
##  1st Qu.:125.2   1st Qu.:20.205   1st Qu.: 8.2798   1st Qu.: 6.5137
##  Median :249.5   Median :35.824   Median :15.8146   Median :10.3853
##  Mean   :249.5   Mean   :37.788   Mean   :15.4350   Mean   :10.1745
##  3rd Qu.:373.8   3rd Qu.:57.853   3rd Qu.:22.5689   3rd Qu.:13.7240
##  Max.   :498.0   Max.   :74.907   Max.   :29.9467   Max.   :20.4535
```

Histograms and Plots

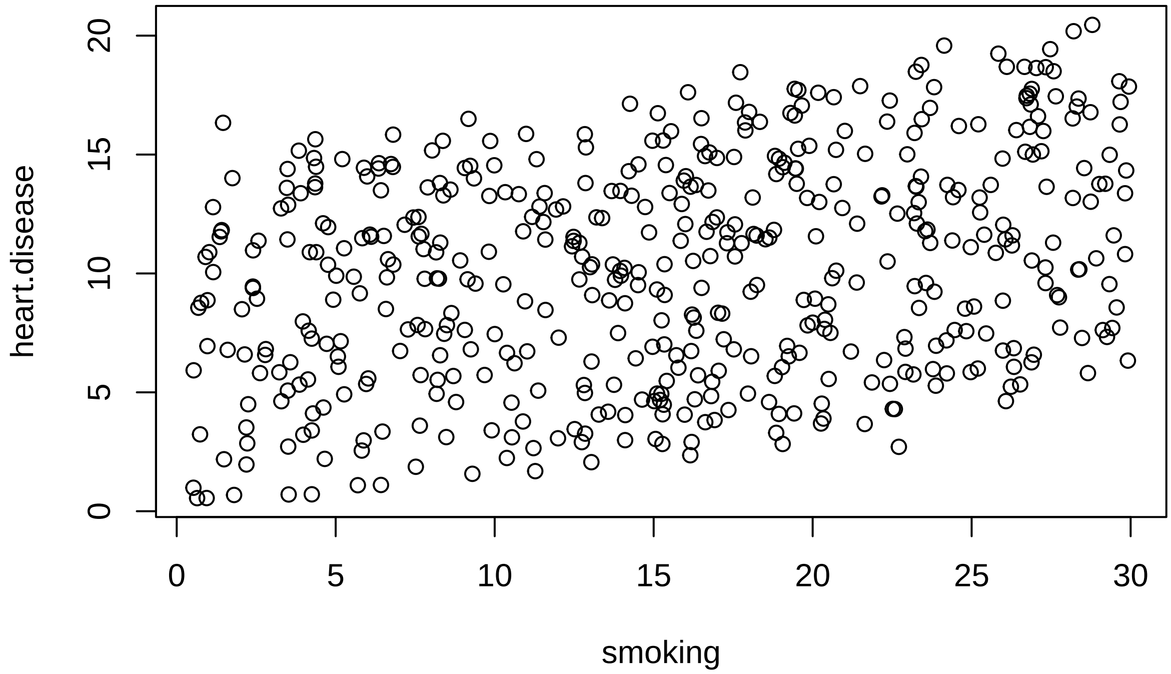
```
hist(heart.data$heart.disease)
```



```r
plot(heart.disease ~ biking, data = heart.data)
```



```r
plot(heart.disease ~ smoking, data = heart.data)
```



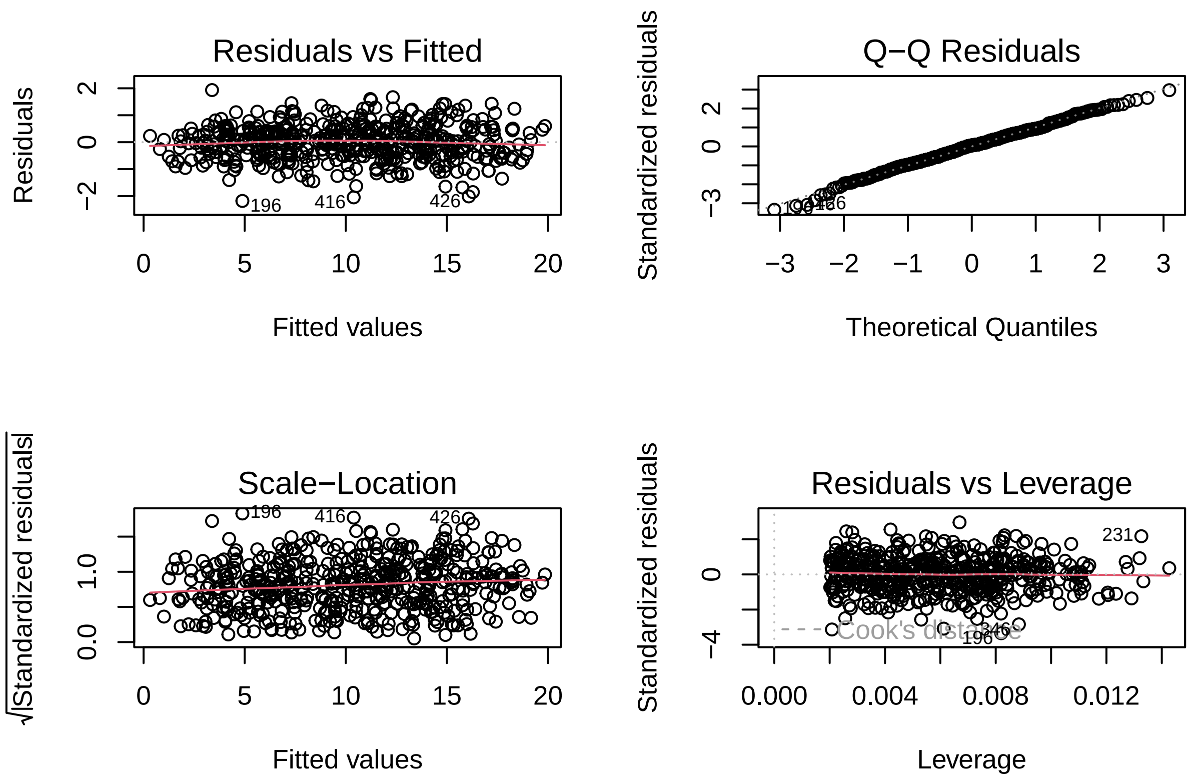
Multiple Regression Model

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

```
##
## Call:
## lm(formula = heart.disease ~ biking + smoking, data = heart.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
```

Diagnostic Plots

```
par(mfrow = c(2, 2))
plot(heart.disease.lm)
```



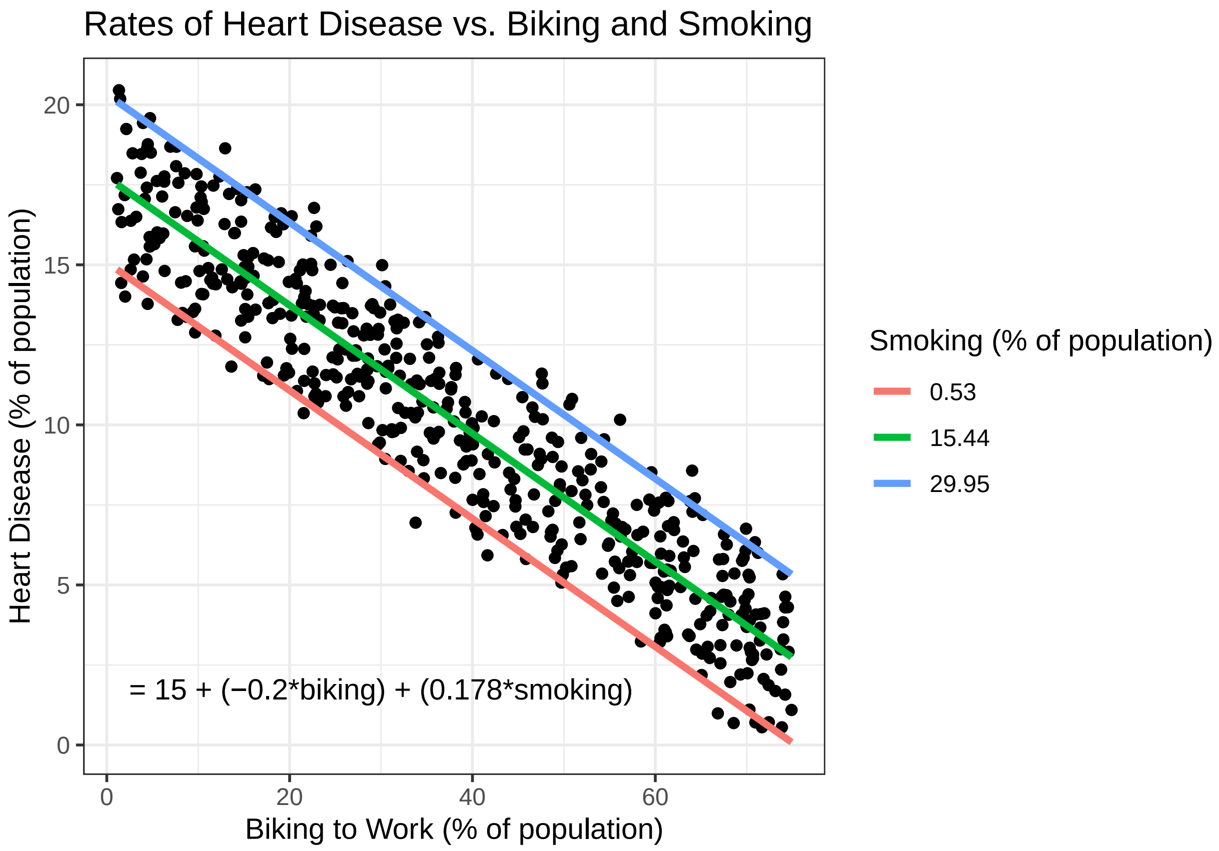
```
par(mfrow = c(1, 1))
```

Plot Predicted Heart Disease Rates

```
plotting.data <- expand.grid(
  biking = seq(min(heart.data$biking), max(heart.data$biking), length.out = 30),
  smoking = c(min(heart.data$smoking), mean(heart.data$smoking), max(heart.data$smoking))
)
plotting.data$predicted.y <- predict(heart.disease.lm, newdata = plotting.data)
plotting.data$smoking <- round(plotting.data$smoking, 2)
plotting.data$smoking <- as.factor(plotting.data$smoking)

heart.plot <- ggplot(heart.data, aes(x = biking, y = heart.disease)) +
  geom_point() +
  geom_line(data = plotting.data, aes(x = biking, y = predicted.y, color = smoking), linewidth = 1.25) +
  theme_bw() +
  labs(
    title = "Rates of Heart Disease vs. Biking and Smoking",
    x = "Biking to Work (% of population)",
    y = "Heart Disease (% of population)",
    color = "Smoking (% of population)"
  ) +
  annotate(geom = "text", x = 30, y = 1.75, label = "= 15 + (-0.2*biking) + (0.178*smoking)")

heart.plot
```



Interpretation

The multiple regression model reveals that increased biking is associated with a significant reduction in heart disease rates, while higher smoking rates are associated with an increase. The model explains 97.96% of the variance in heart disease prevalence, indicating a very strong fit.