RM_R_code

2025-04-17

Aakansha Rawat

Getting and setting the directory

```
getwd()
```

```
## [1] "C:/Users/Hamada Salim G Trd/Desktop/coding_Or_progrmming"
```

```
setwd("C://Users//Hamada Salim G Trd//Desktop//coding_Or_progrmming")
getwd()
```

```
## [1] "C:/Users/Hamada Salim G Trd/Desktop/coding_Or_progrmming"
```

Retrieving the file

```
library(readxl)
data1<-read_excel("real_RM.xlsx")
summary(data1)
```

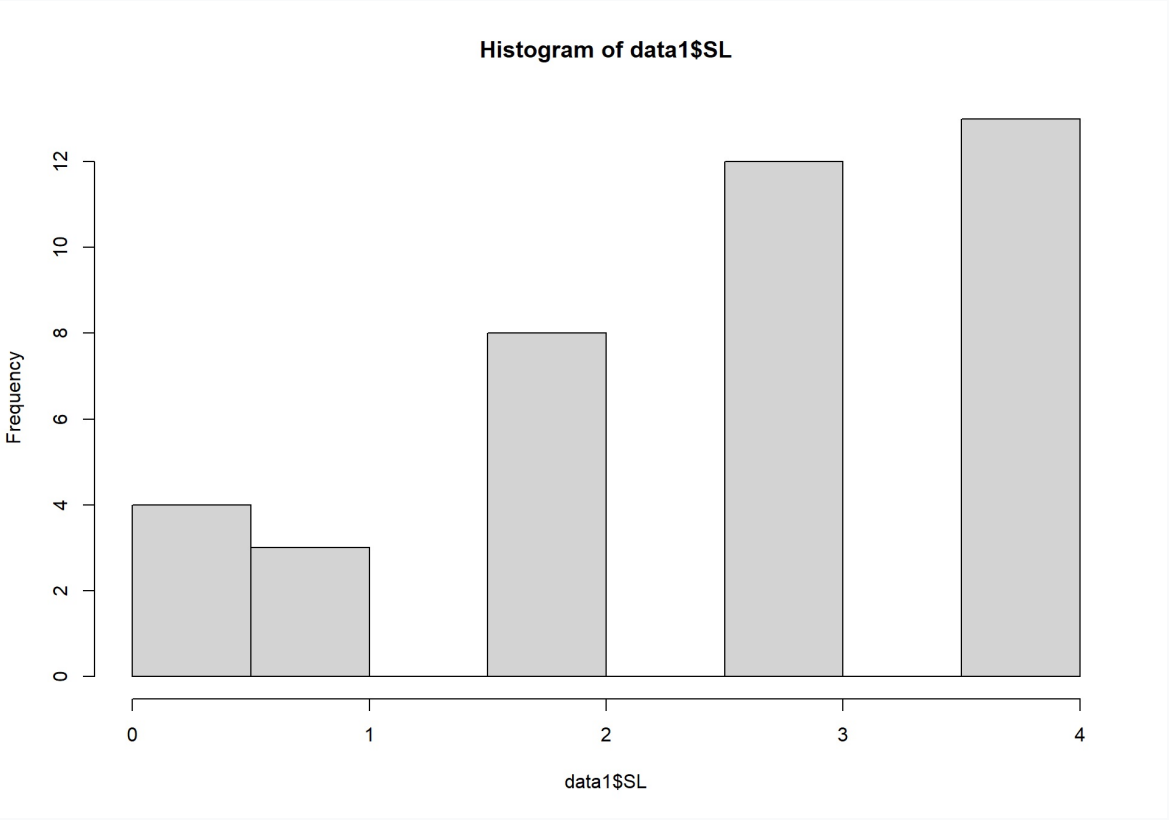
```
##       AGE           OCCUP              SCOE (Y)         WOE (M)
##  Min.   :19.00   Length:40          Min.   : 15000   Min.   :10000
```

```
##  1st Qu.:20.00   Class :character   1st Qu.: 15000   1st Qu.:20000
##  Median :20.00   Mode  :character   Median : 45000   Median :30000
##  Mean   :22.62                      Mean   : 68919   Mean   :43333
##  3rd Qu.:21.00                      3rd Qu.:105000   3rd Qu.:60000
##  Max.   :51.00                      Max.   :150000   Max.   :90000
##                                     NA's   :3        NA's   :37
##     WOE (Y)           STATE                NTVO         HGO
##  Min.   : 120000   Length:40          Min.   :1.0   Length:40
##  1st Qu.: 240000   Class :character   1st Qu.:1.0   Class :character
##  Median : 360000   Mode  :character   Median :3.0   Mode  :character
##  Mean   : 520000                      Mean   :2.7
##  3rd Qu.: 720000                      3rd Qu.:4.0
##  Max.   :1080000                      Max.   :5.0
##  NA's   :37
##     TTR (M)             SL           TWL (M)         TWL (H)
##  Min.   : 15.00   Min.   :0.000   Min.   : 15.0   Min.   :0.250
##  1st Qu.: 15.00   1st Qu.:2.000   1st Qu.: 15.0   1st Qu.:0.250
##  Median : 45.00   Median :3.000   Median : 67.5   Median :1.125
##  Mean   : 43.88   Mean   :2.675   Mean   : 75.0   Mean   :1.250
##  3rd Qu.: 52.50   3rd Qu.:4.000   3rd Qu.:150.0   3rd Qu.:2.500
##  Max.   :150.00   Max.   :4.000   Max.   :150.0   Max.   :2.500
##
##      MPW               WDWWL                MSWL             HOGS
##  Length:40          Length:40          Min.   :  0.00   Min.   :1.00
##  Class :character   Class :character   1st Qu.: 25.00   1st Qu.:2.00
##  Mode  :character   Mode  :character   Median : 75.00   Median :3.00
##                                        Mean   : 83.75   Mean   :2.75
##                                        3rd Qu.:150.00   3rd Qu.:3.00
##                                        Max.   :250.00   Max.   :5.00
##
##       PPU              EF             HSC              HEW
##  Min.   :1.000   Min.   :1.000   Min.   :  0.0   Min.   :  0.0
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 10.0   1st Qu.:  0.0
##  Median :3.000   Median :3.000   Median : 30.0   Median :  0.0
##  Mean   :2.775   Mean   :2.763   Mean   : 42.5   Mean   : 19.5
##  3rd Qu.:3.250   3rd Qu.:3.500   3rd Qu.: 70.0   3rd Qu.:  0.0
##  Max.   :5.000   Max.   :5.000   Max.   :100.0   Max.   :540.0
##
##     TTW (H)           OCTW             TTC              TML
##  Min.   :0.750   Min.   :   7.5   Min.   :  5.00   Min.   :  17.5
##  1st Qu.:1.250   1st Qu.:  47.5   1st Qu.: 22.50   1st Qu.: 233.1
##  Median :2.875   Median :  90.0   Median : 45.00   Median : 540.0
##  Mean   :2.712   Mean   : 152.9   Mean   : 85.62   Mean   : 755.2
##  3rd Qu.:3.438   3rd Qu.: 202.5   3rd Qu.:110.00   3rd Qu.:1202.5
##  Max.   :7.500   Max.   :1080.0   Max.   :630.00   Max.   :2250.0
##
```

```
names(data1)
```

```
##  [1] "AGE"      "OCCUP"    "SCOE (Y)" "WOE (M)"  "WOE (Y)"  "STATE"
##  [7] "NTVO"     "HGO"      "TTR (M)"  "SL"       "TWL (M)"  "TWL (H)"
## [13] "MPW"      "WDWWL"    "MSWL"     "HOGS"     "PPU"      "EF"
## [19] "HSC"      "HEW"      "TTW (H)"  "OCTW"     "TTC"      "TML"
```

```
hist(data1$SL)
```




```
table(data1$SL)
```

```
##
##  0  1  2  3  4
##  4  3  8 12 13
```

Stress levels were unevenly distributed, suggesting respondents perceived most situations as moderately-to-highly stressful regardless of time/money costs.

Regression model sample 1

```
model1<-lm(SL~ `TWL (H)` + TML + EF,data = data1)
model1
```

```
##
## Call:
## lm(formula = SL ~ `TWL (H)` + TML + EF, data = data1)
##
## Coefficients:
## (Intercept)    `TWL (H)`          TML           EF
##   1.6580229    0.4179732    0.0006134    0.0113104
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = SL ~ `TWL (H)` + TML + EF, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.96336 -0.53783  0.00571  0.89808  1.96222
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.6580229  0.8111030   2.044   0.0483 *
## `TWL (H)`   0.4179732  0.2255956   1.853   0.0721 .
## TML         0.0006134  0.0003262   1.881   0.0681 .
## EF          0.0113104  0.2119797   0.053   0.9577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.183 on 36 degrees of freedom
## Multiple R-squared:  0.2228, Adjusted R-squared:  0.1581
## F-statistic: 3.441 on 3 and 36 DF,  p-value: 0.02677
```

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: SL
##           Df Sum Sq Mean Sq F value  Pr(>F)
## `TWL (H)`  1  9.102  9.1024  6.5094 0.01512 *
## TML        1  5.329  5.3286  3.8107 0.05874 .
## EF         1  0.004  0.0040  0.0028 0.95774
## Residuals 36 50.340  1.3983
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

- Now i have first formed a linear regression model named "model1" with Y = B_1 + B_2X_2 + B_3X_3 + B_4X_4 + u *as the* ***PRF*** *and* $Y\ hat = b_1 + b_2X_2 + b_3X_3 + b_4X_4$ + e is the **SRF** where Y is the stress level and X1 is the total waiting time in line, X2 is total monetary loss and X3 is efficiency.
- So the summary shows us that $\mathbf{b_1 = 1.6580229}$, $\mathbf{b_2 = 0.4179732}$,$\mathbf{b_3 = 0.0006134}$ and $\mathbf{b_4 = 0.0113104}$ we also have residual standard error i.e sigma hat which is **1.1825119** also called as *standard error of regression* (**SER**).
- r^2 value is **0.2228** i.e only 22.28% which means only this percentage of variation in Y was explained by the model also known as **goodness of fit.**
- The density plot of residuals shows us that they are not perfectly normally distributed but slightly positively skewed.
- There is the **adjusted R^2** *value* which is **0.1580851**

```
pv1<-vcov(model1)
pv1
```

```
##               (Intercept)     `TWL (H)`           TML            EF
## (Intercept)  0.6578881222 -8.544387e-02 -1.194003e-04 -1.541893e-01
## `TWL (H)`   -0.0854438705  5.089339e-02 -1.389908e-05  1.170114e-02
## TML         -0.0001194003 -1.389908e-05  1.063749e-07  2.042878e-05
## EF          -0.1541892761  1.170114e-02  2.042878e-05  4.493539e-02
```

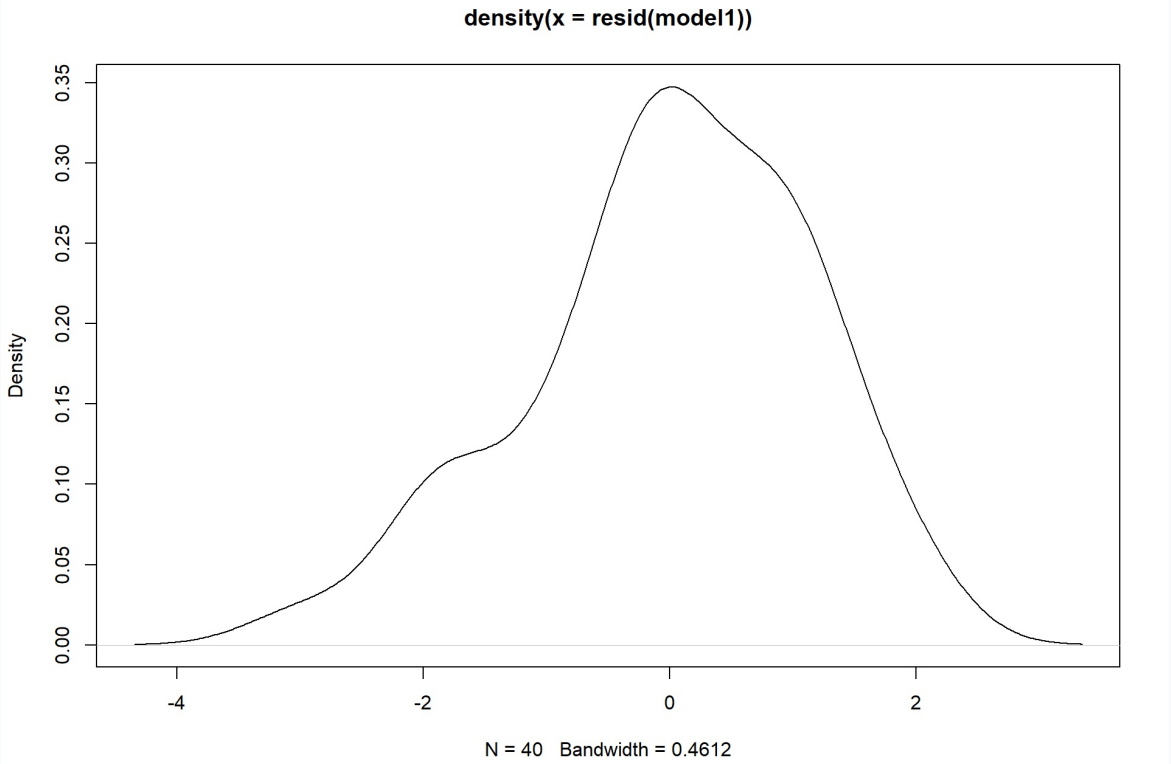
```
variances1 <- diag(pv1)
variances1
```

```
##  (Intercept)    `TWL (H)`          TML           EF
## 6.578881e-01 5.089339e-02 1.063749e-07 4.493539e-02
```

```
resid1 <- resid(model1)
var(resid1)
```

```
## [1] 1.29077
```

```
plot(density(resid(model1)))
```



- vcov() function is used to extract the variance-covariance matrix in which the diagonal elements of this matrix represent the variances of the parameter estimates. Similary, we find the variance of residuals.
- the density plot shows that it is slightly positively skewed and notnormally distributed as one of assumptions of CLRM.

Regression model sample 2

```
data1$MSWL_100 <- data1$MSWL / 100
model2 <- lm(SL ~ `TWL (H)` + MSWL_100, data = data1)
model2
```

```
##
## Call:
## lm(formula = SL ~ `TWL (H)` + MSWL_100, data = data1)
##
## Coefficients:
## (Intercept)    `TWL (H)`     MSWL_100
```

```
##      1.6424       0.3624       0.6920
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = SL ~ `TWL (H)` + MSWL_100, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0675 -0.7051  0.2810  0.7550  1.9128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6424     0.3227   5.090 1.07e-05 ***
## `TWL (H)`     0.3624     0.2033   1.783  0.08279 .
## MSWL_100      0.6920     0.2293   3.017  0.00459 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.099 on 37 degrees of freedom
## Multiple R-squared:  0.3102, Adjusted R-squared:  0.273
## F-statistic: 8.321 on 2 and 37 DF,  p-value: 0.001037
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: SL
##           Df Sum Sq Mean Sq F value   Pr(>F)
## `TWL (H)`  1  9.102  9.1024  7.5380 0.009270 **
## MSWL_100   1 10.994 10.9939  9.1044 0.004595 **
## Residuals 37 44.679  1.2075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

- Now i have first formed a linear regression model named "sm" with Y = B_1 + B_2X_2 + B_3X_3 + u as the **PRF** and Y hat = b_1 + b_2X_2 + b_3X_3 + e is the **SRF** where Y is the stress level and X1 is the total time wasted in line and X2 is the

money spent while waiting in line.

- So the summary shows us that $\mathbf{b_1 = 1.6424}$, $\mathbf{b_2 = 0.3624}$ and $\mathbf{b_3 = 0.6920}$ we also have residual standard error i.e sigma hat which is **1.0988785**
- r^2 value is **0.3102** i.e only 31.02% which means only this percentage of variation in Y was explained by the model also known as **goodness of fit.**
- vcov() function is used to extract the variance-covariance matrix in which the diagonal elements of this matrix represent the variances of the parameter estimates. Similary, we find the variance of residuals.
- The density plot of residuals shows us that they are not perfectly normally distributed but slightly positively skewed.
- There is the **adjusted $\mathbf{R}^2$** *value* which is **0.272963**

```
pv2<-vcov(model2)
pv2
```

```
##             (Intercept)   `TWL (H)`    MSWL_100
## (Intercept)  0.10411056 -0.04063769 -0.02761205
## `TWL (H)`   -0.04063769  0.04131971 -0.01314859
## MSWL_100    -0.02761205 -0.01314859  0.05259437
```

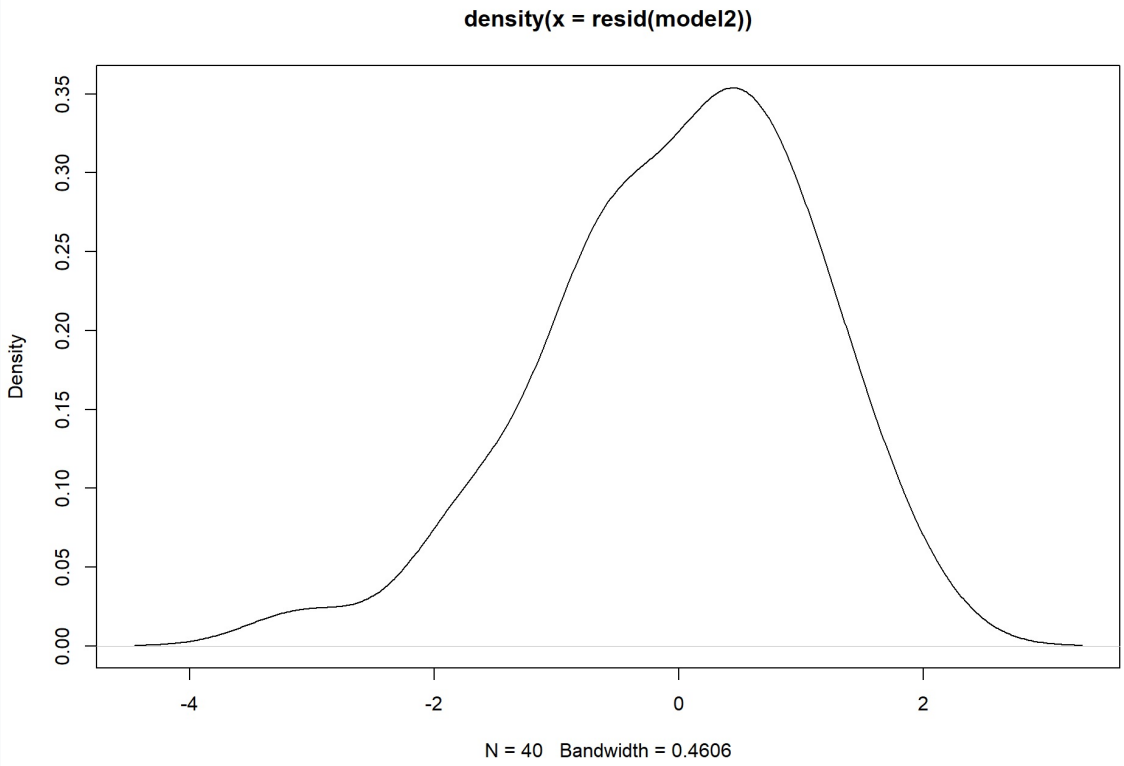
```
variances2 <- diag(pv2)
variances2
```

```
## (Intercept)   `TWL (H)`    MSWL_100
##  0.10411056  0.04131971  0.05259437
```

```
resid2 <- resid(model2)
var(resid2)
```

```
## [1] 1.145609
```

```
plot(density(resid(model2)))
```



- vcov() function is used to extract the variance-covariance matrix in which the diagonal elements of this matrix represent the variances of the parameter estimates. and again this is also slightly positively skewed.

Regression model sample 3

```
data1$MPW <- ifelse(data1$MPW == "Yes", 1, 0)
model3 <- lm(SL ~ `TWL (H)` + TML + MPW, data = data1)
summary(model3)
```

```
##
## Call:
## lm(formula = SL ~ `TWL (H)` + TML + MPW, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.61606 -0.63949  0.03439  0.79272  2.19801
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.5091681  0.4162342   3.626 0.000884 ***
## `TWL (H)`   0.3653891  0.2239817   1.631 0.111539
## TML         0.0005373  0.0003191   1.684 0.100928
## MPW         0.4044257  0.4674726   0.865 0.392695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.17 on 36 degrees of freedom
## Multiple R-squared:  0.2386, Adjusted R-squared:  0.1752
## F-statistic: 3.761 on 3 and 36 DF,  p-value: 0.01903
```

For Heteroscadasticity

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 2.8657, df = 3, p-value = 0.4128
```

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 2.2478, df = 2, p-value = 0.325
```

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 9.6432, df = 3, p-value = 0.02186
```

To check the presence of heteroskedasticity in the regression models, the studentized Breusch-Pagan test was performed. For Model 1 and Model 2, the p-values were 0.4128 and 0.325 respectively, indicating that the null hypothesis of homoskedasticity (constant variance of residuals) cannot be rejected at conventional significance levels. This suggests that these models do not suffer from heteroskedasticity. However, in Model 3, the test produced a p-value of 0.02186, which is statistically significant at the 5% level. This indicates the presence of heteroskedasticity in Model 3, implying

that the assumption of constant variance of the residuals is violated. This could affect the efficiency of the estimators in Model 3, and caution should be taken while interpreting its results.

- The Breusch-Pagan test for Model 3 returned a **p-value of 0.02186**, indicating the presence of **heteroskedasticity** at the 5% level. This violates the assumption of constant error variance in OLS regression and can lead to unreliable standard errors and incorrect significance testing.

To address this issue, **robust standard errors** were applied using the **heteroskedasticity-consistent covariance matrix estimator (HC1)**.

```
coeftest(model3, vcov = vcovHC(model3, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.50916810 0.53357916  2.8284 0.007597 **
## `TWL (H)`   0.36538910 0.28373409  1.2878 0.206037
## TML         0.00053727 0.00028628  1.8768 0.068673 .
## MPW         0.40442575 0.63348640  0.6384 0.527246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

While the coefficients remained unchanged, the significance of monetary loss (TML) strengthened to marginal levels (p=0.069), whereas time spent waiting (TWL_H) lost significance. This suggests the original standard errors understated TML's precision and overstated TWL_H's importance. This improves the reliability of the inference made from Model 3.

For Multicollinearity

```
vif(model1)
```

```
## `TWL (H)`       TML        EF
##  1.155564  1.190305  1.220927
```

```
vif(model2)
```

```
## `TWL (H)`  MSWL_100
##   1.08643   1.08643
```

```
vif(model3)
```

```
## `TWL (H)`       TML       MPW
##  1.162679  1.163259  1.196368
```

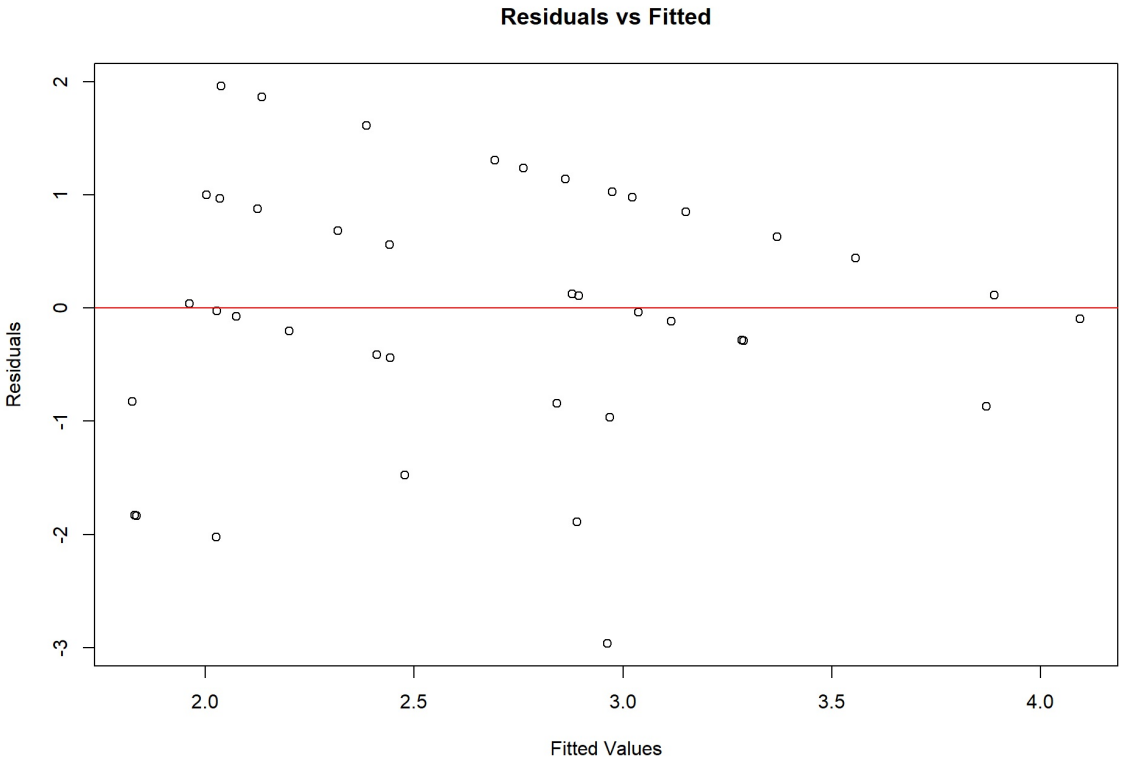
To assess multicollinearity among the independent variables, Variance Inflation Factor (VIF) values were computed for all three models. For Model 1, the VIFs were as follows: TWL (H) = 1.16, TML = 1.19, and EF = 1.22. For Model 2, TWL (H) and MSWL_100 both had a VIF of 1.09. In Model 3, the VIF values were: TWL (H) = 1.16, TML = 1.16, and MPW = 1.20.

Since all VIF values are well below the commonly accepted threshold of 5 (or 10), there is no evidence of serious multicollinearity among the predictor variables in any of the models. This suggests that the estimated coefficients are stable and reliable, and multicollinearity is not a concern in this analysis.

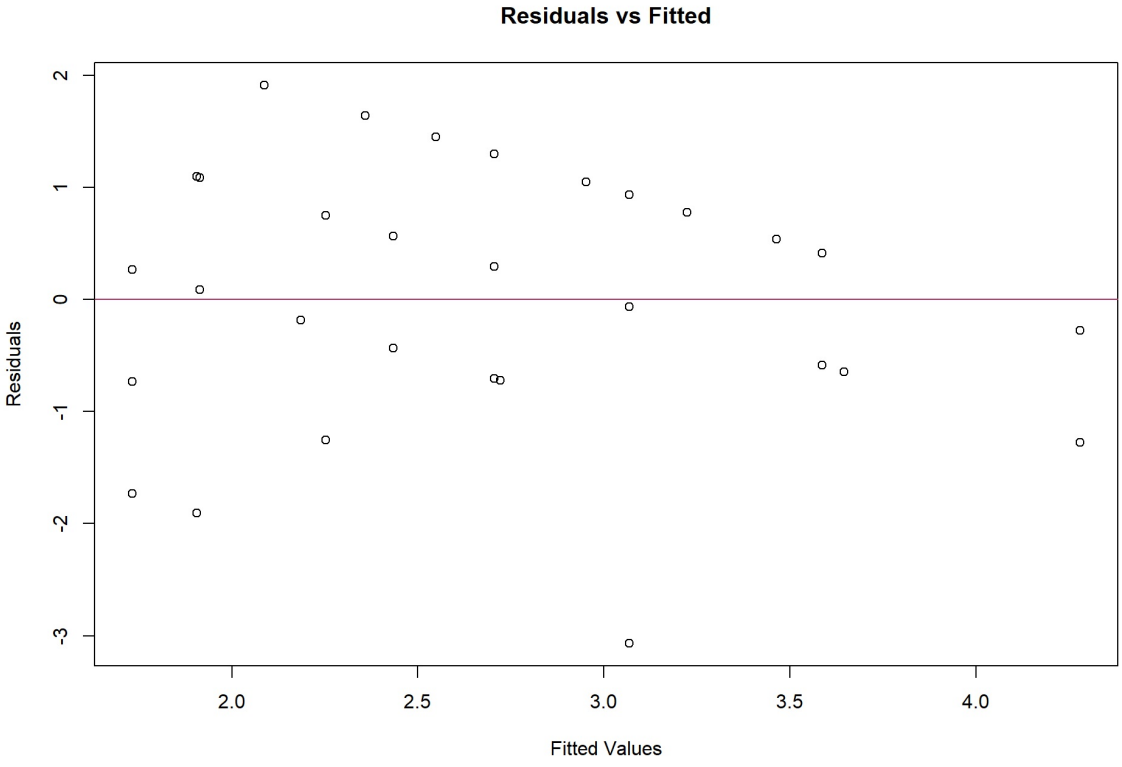
Model 2 shows the lowest VIFs, supporting its use as the most parsimonious specification.

For Autocorrelation

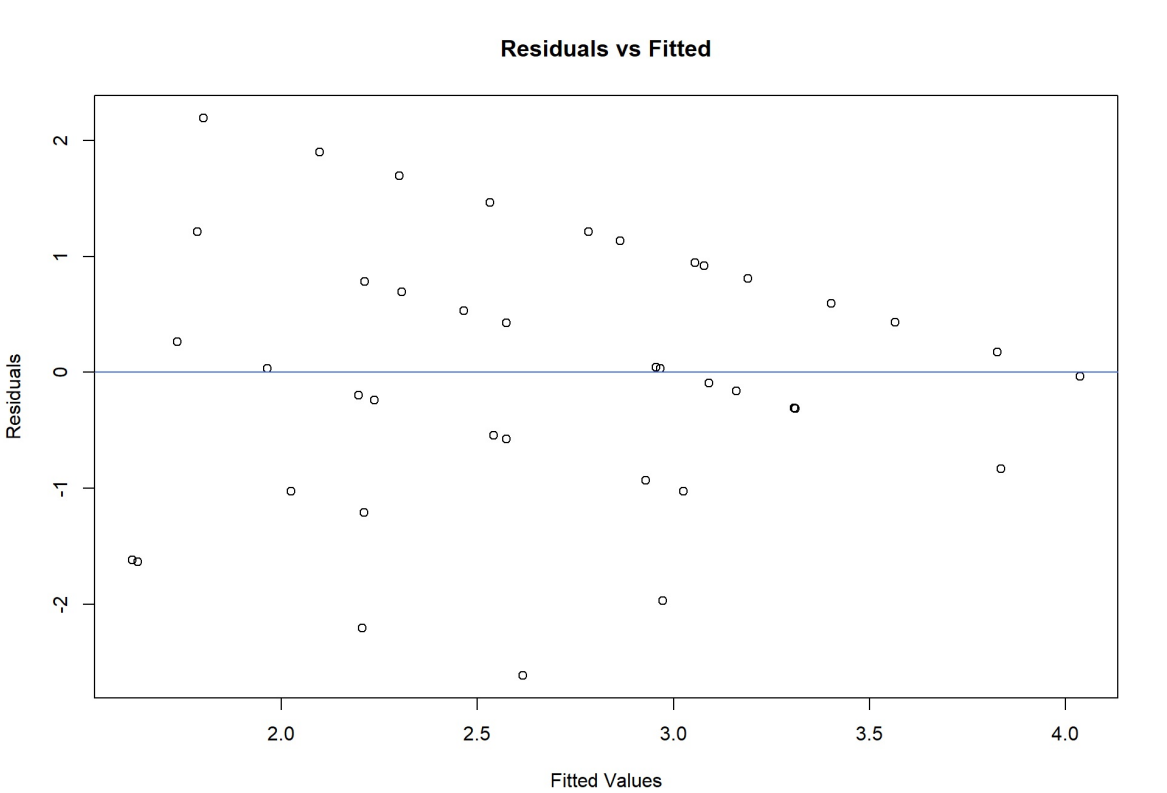
```
plot(model1$fitted.values, resid(model1),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "red")
```




```r
plot(model2$fitted.values, resid(model2),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "maroon")
```




```r
plot(model3$fitted.values, resid(model3),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "royalblue")
```




```
dwtest(model1)
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 2.2372, p-value = 0.7798
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(model2)
```

```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 2.0655, p-value = 0.584
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(model3)
```

```
##
```

```
##  Durbin-Watson test
##
## data:  model3
## DW = 2.2756, p-value = 0.8201
## alternative hypothesis: true autocorrelation is greater than 0
```

To check for autocorrelation in the residuals of the regression models, the Durbin-Watson (DW) test was performed for each model. The DW statistic for Model 1 was 2.24 ($p = 0.7798$), for Model 2 it was 2.07 ($p = 0.584$), and for Model 3, the value was 2.28 ($p = 0.8201$).

Since all the DW statistics are close to 2, and the p-values are not statistically significant, we fail to reject the null hypothesis of no autocorrelation. This indicates that there is no evidence of positive autocorrelation in the residuals of any of the models, confirming the assumption of independently distributed errors.

Checking correlation

```
cor(data1$SL,data1$EF)
```

```
## [1] -0.1932616
```

```
cor(data1$`TWL (H)`,data1$SL)
```

```
## [1] 0.3748637
```

```
cor(data1$SL,data1$TML)
```

```
## [1] 0.3808814
```

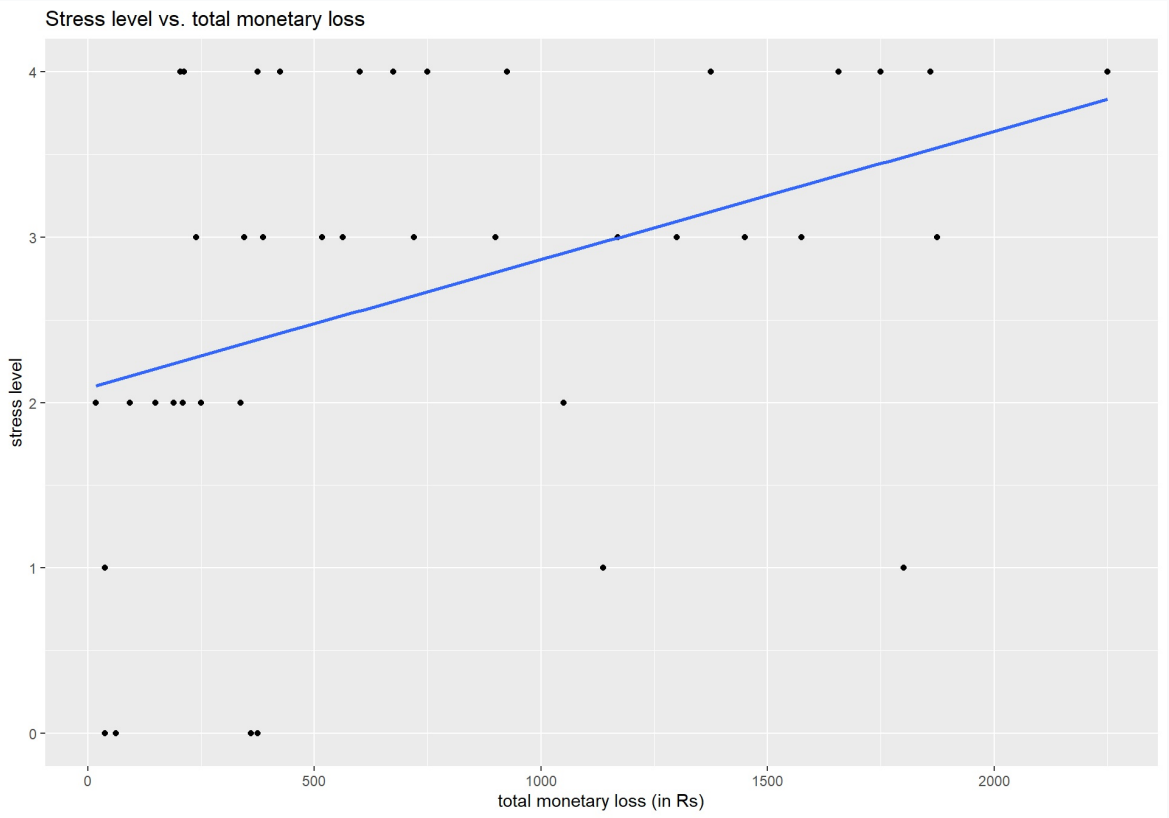
Now here I have calculated the correlation of each explanatory variable with the dependent variable which is **-0.1932616**, **0.3748637**, and **0.3808814**. We see there is a -ve relationship between stress level and efficiency, whereas a +ve relationship between stress level and total waiting time in line and total monetary loss, though in last 2 cases there is a *moderate correlation* and *weak negative linear relation* in the 1st case, which is also very evident from the corresponding graphs.

Plotting

```r
plot1 <- ggplot(model1, aes(x = TML, y = SL)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Stress level vs. total monetary loss",
       x = "total monetary loss (in Rs)",
       y = "stress level")

plot1
```

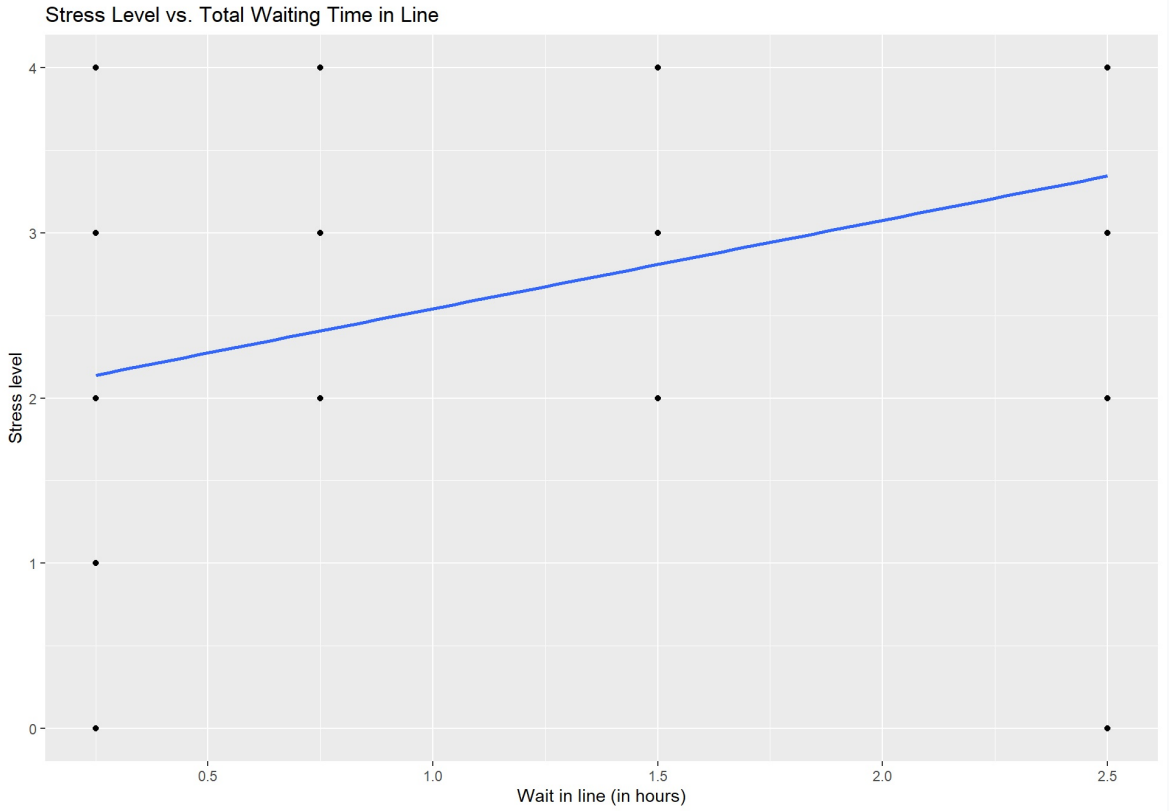
```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
plot2 <- ggplot(data1, aes(x = `TWL (H)`, y = SL)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Stress Level vs. Total Waiting Time in Line",
    x = "Wait in line (in hours)",
    y = "Stress level"
  )

print(plot2)
```

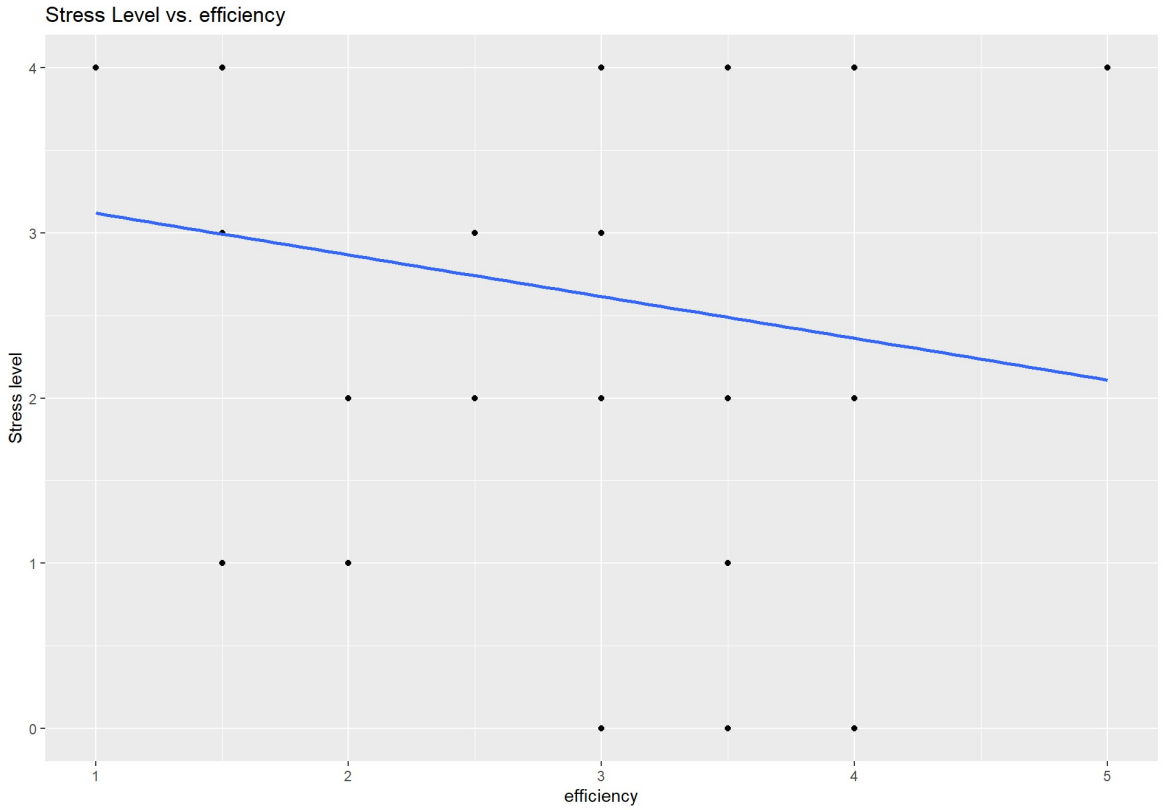
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
plot3 <- ggplot(data1, aes(x = EF, y = SL)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Stress Level vs. efficiency",
    x = "efficiency",
    y = "Stress level"
  )

print(plot3)
```

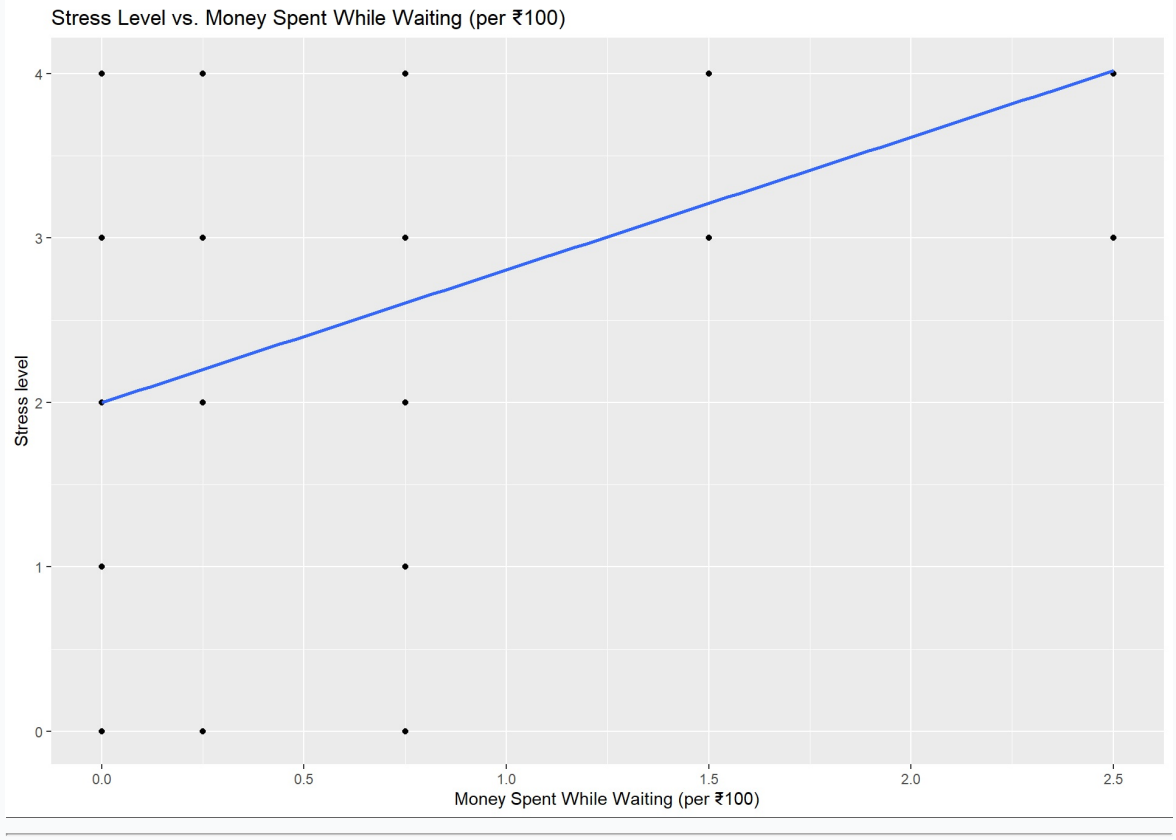
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
plot4 <- ggplot(data1, aes(x = MSWL_100, y = SL)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Stress Level vs. Money Spent While Waiting (per ₹100)",
    x = "Money Spent While Waiting (per ₹100)",
    y = "Stress level"
  )

print(plot4)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



1. *SER represents the standard deviation of the errors/residuals in a regression model. It measures the accuracy of the regression predictions.*↩

2. r^2 is a statistical measure used in regression analysis to assess the goodness of fit of a regression model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables.↩

3. is a modification of the standard r^2 value that adjusts for the number of predictors in a regression model. While r^2 tends to increase as more predictors are added to the model, even if those predictors are not relevant, Adjusted R^2 penalizes the addition of unnecessary variables.↩