MAS Workshop

Department of Statistics - USJ

2024-03-28

```
# load the packages

library(readxl)   # to load the data set
library(dplyr)    # select operator
library(ggplot2)  # to create plots
library(magrittr) # pipe operator
library(car)      # to obtain vif value
```

Multiple linear regression analysis

Example:
Suppose we aim to identify the factors affecting earnings from the product sales in the apparel industry. To examine the relationship between selected variables and earnings, we will conduct a multiple regression analysis. For this analysis, we will utilize the variables Earnings, MOH Value, and Std Hrs. The response variable is earnings whereas MOH Value and Std Hrs are the predictor variables.

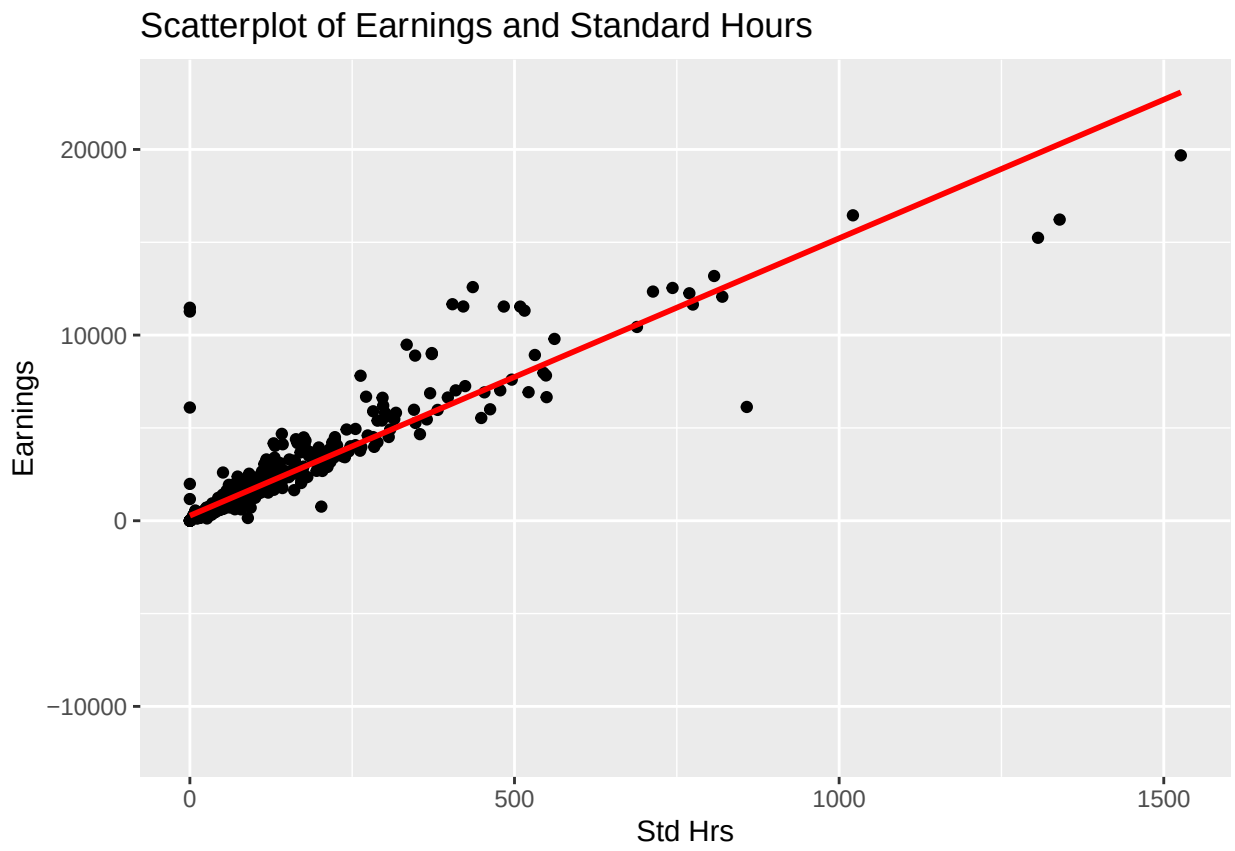
```
# load the data set
MAS_data_set <- read_excel("data set 2.xlsx")

# create a data set for regression analysis
Reg_data <- MAS_data_set %>%
  select(`MOH Value`, `Earnings`,`Std Hrs`)
```

Check the linearity assumption

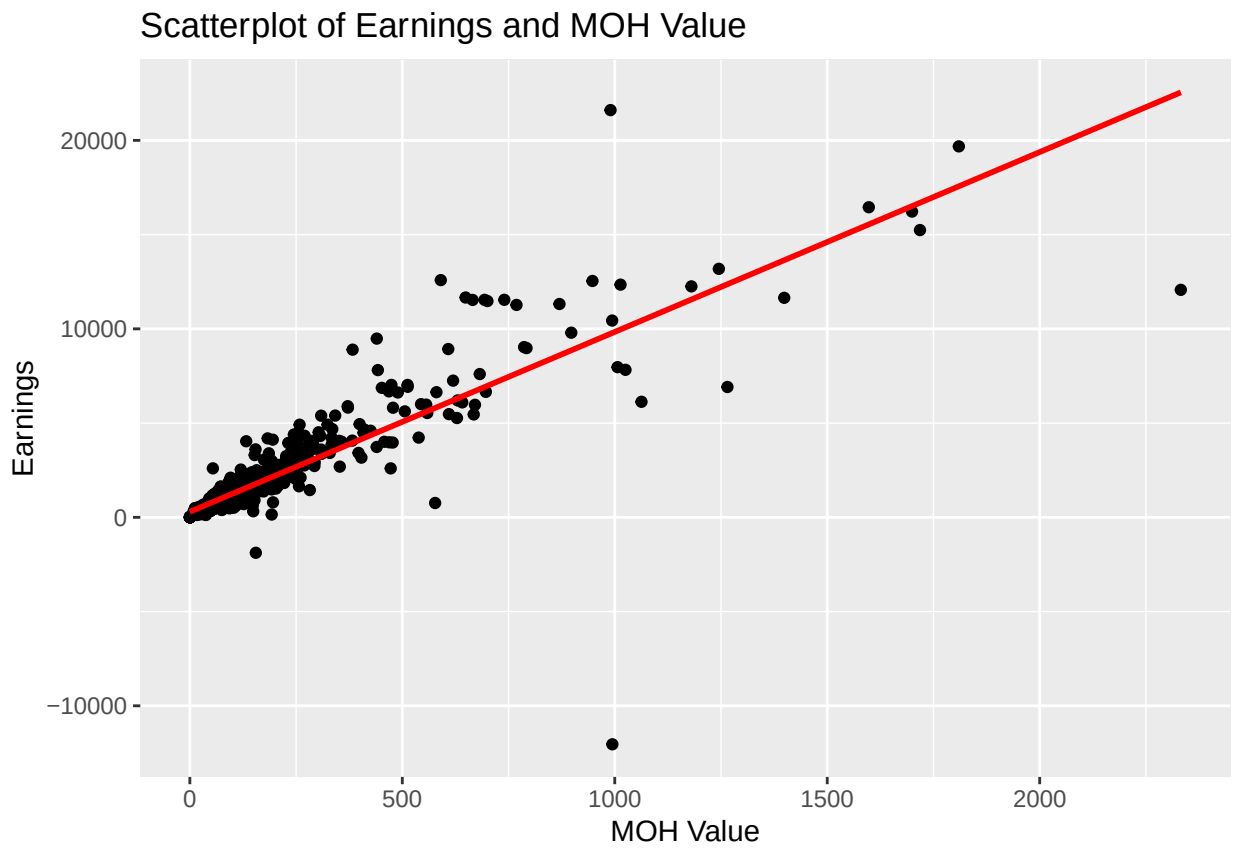
```
# scatter plot of Earnings and Standard Hours

ggplot(Reg_data, aes(x=`Std Hrs`, y=Earnings)) +
  geom_point() + geom_smooth(method = lm, se = FALSE, color = "red")+
  ggtitle("Scatterplot of Earnings and Standard Hours")
```




```
# scatter plot of Earnings and MOH Value

ggplot(Reg_data, aes(x=`MOH Value`, y=Earnings)) +
  geom_point() + geom_smooth(method = lm, se = FALSE, color = "red")+
  ggtitle("Scatterplot of Earnings and MOH Value")
```



Fit the model

```
##
## Call:
## lm(formula = Earnings ~ `Std Hrs` + `MOH Value`, data = Reg_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6847.3  -248.4  -157.2    60.3  7315.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 225.9173    44.2228   5.109  4.3e-07 ***
## `Std Hrs`     6.8198     0.6583  10.359  < 2e-16 ***
## `MOH Value`   5.6160     0.4265  13.167  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 934.3 on 636 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.8733, Adjusted R-squared:  0.8729
## F-statistic:  2192 on 2 and 636 DF,  p-value: < 2.2e-16
```

Assumption checking

Check the multicollinearity assumption

```
vif(model_reg)  # Since VIF values are less than 10, we can avoid the multicollinearity
```

```
##   `Std Hrs` `MOH Value`
##    8.174816    8.174816
```

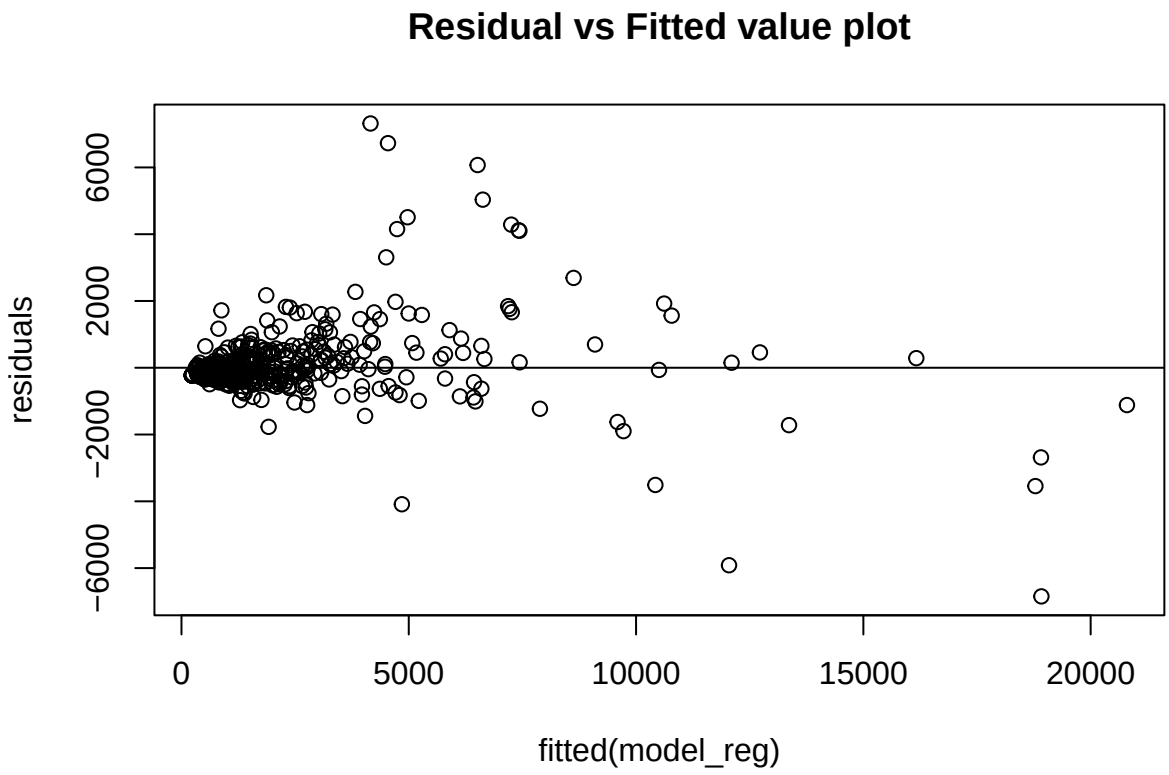
Check the constant variance assumption of residuals

```
# obtain the residuals
residuals <- resid(model_reg)

# residual vs. fitted value plot
plot(fitted(model_reg), residuals) + title("Residual vs Fitted value plot")
```

```
## integer(0)
```

```
# add a horizontal line at 0
abline(0,0)
```



Check the normal assumption of residuals

```
# Q-Q plot for residuals
qqnorm(residuals)

# add a straight diagonal line to the plot
qqline(residuals)
```