

R Assignment 5

Jasmeet Singh Saini - 0758054

2023-04-11

Question 1 - Local Advertising

Required Data: *advertising.csv*

Upon request, the local newspaper will provide data to help potential advertisers target their ads. The paper provides a random sample of 12 advertisements per day, 4 from each of the three sections of the paper, for the previous week and the number of inquiries the ad generated for the business. Use *advertising.csv* to answer the following questions.

Firstly, let's load the data set and check the structure of *advertising.csv*.

```
# Importing the dataset
advertising_data <- read.csv("advertising.csv")
# Let's view the str()
str(advertising_data)
```

```
## 'data.frame':    60 obs. of  3 variables:
## $ Day          : chr  "Monday" "Monday" "Monday" "Monday" ...
## $ Section      : chr  "News" "News" "News" "News" ...
## $ Inquiries    : int   11  8  6  8  9 10 10 12  8  9 ...
```

Here, we can see that “advertising_data” is in long format and the categorical data variables are of “character” data type. Only Inquiries is numeric variable and its data type is integer.

In this part of question, we will assume α , level of significance as **0.01**.

a) Ignoring any potential interaction between Day and Section, are there any differences in the average number of inquiries by day? (Note: this is a hypothesis test and should be structured as such.)

We will use **One-Way ANOVA Test** to check whether there are any differences in the average number of inquiries by day.

Step 1 : Hypothesis & Assumptions

Step 1a: Hypothesis The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no significant difference in the average number of inquiries by day.

vs.

H_A : There is significant difference in the average number of inquiries by day

Step 1b: Assumptions

We need to check the assumptions of:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Normality** : The residuals should be nearly normal distributed and show normality. It will be checked by fitting the linear model and using Q-Q Plot and Shapiro-Wilk's Test of model residuals.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals). It will be checked by using Residual vs Fitted plot of model residuals and Levene's Test

Before we fitting the `lm()` model and check the the assumptions of Independence, Normality and Variance, we need to convert the categorical variables to factor data type. Here, we will make "Day" variable a ordered factor and "Section" variable a normal factor.

```
advertising_data$Day <- factor(advertising_data$Day
                              , levels = c("Monday","Tuesday","Wednesday","Thursday","Friday")
                              , ordered = TRUE)

advertising_data$Section <- factor(advertising_data$Section)
```

Now let's again check the `str()` of `advertising_data`

```
str(advertising_data)

## 'data.frame':    60 obs. of  3 variables:
## $ Day          : Ord.factor w/ 5 levels "Monday"<"Tuesday"<...: 1 1 1 1 2 2 2 2 3 3 ...
## $ Section      : Factor w/ 3 levels "Business","News",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ Inquiries    : int  11 8 6 8 9 10 10 12 8 9 ...
```

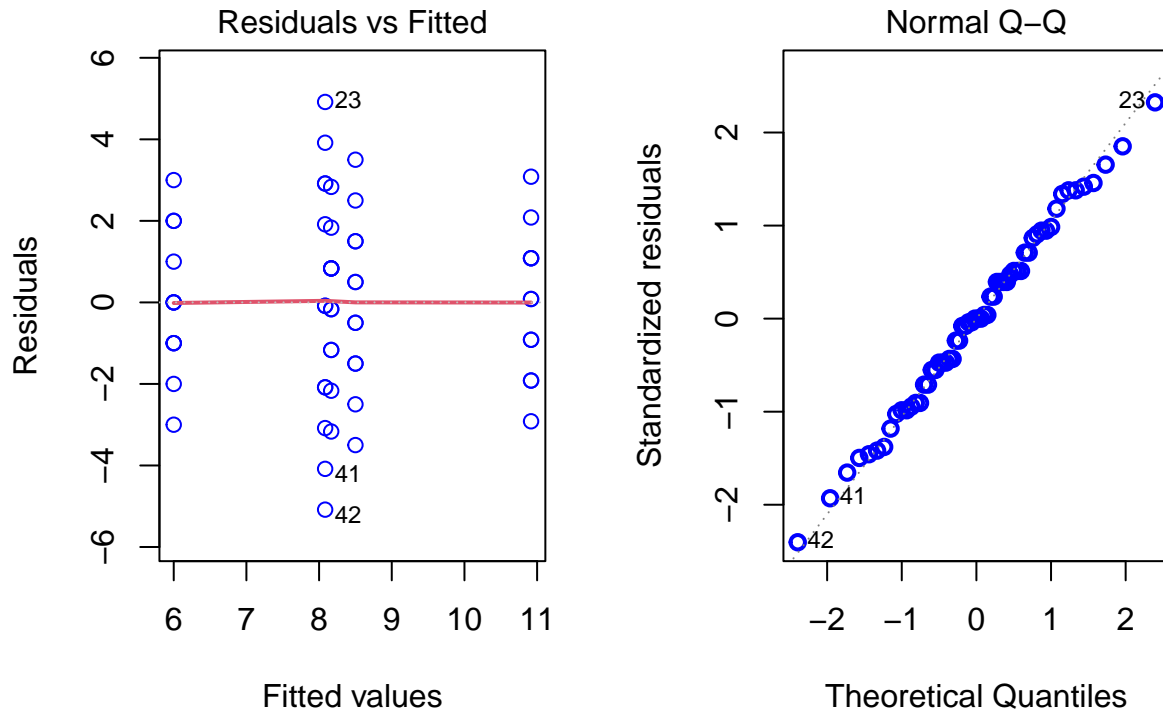
Now, let's fit the linear model:

```
linear_model <- lm(Inquiries ~ Day, data = advertising_data)
```

Checking the **assumptions**:

Plot of Residual vs Fitted Plot & QQ Plot:

```
par(mfrow=c(1,2))
plot(linear_model, which = 1:2, col = "blue", lwd = 2, pch = 1)
```



Histogram of residuals:

```
hist(linear_model$residuals, main = "Distribution of residuals",
     , xlab = "residuals", col = "lightblue")
```



Checking normality using Shapiro-Wilk's Test:

```
shapiro.test(linear_model$residuals)
```

```
##
## Shapiro-Wilk normality test
##
## data: linear_model$residuals
## W = 0.99351, p-value = 0.9876
```

As it can be observed in the Q-Q Plot that the data points are closely attached to the model fitted line and histogram has a frequency peaks in the center. Also, p-value in Shapiro-Wilk normality test is **0.9876**. Hence, $p\text{-value} > 0.01$ and we can say that the data is normal.

```
# Levene's Test for Homogeneity of Variance
car::leveneTest(linear_model)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 4  3.0279 0.02503 *
##      55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the Residual vs Fitted plot, all the data points are in the constant distance to the line and p-value in Levene's test is **0.02503**. Hence, $p\text{-value} > 0.01$ and we can say that constant variance is reached.

Hence, all of the conditions are met and now we can proceed with the testing using the parametric approach.

Step 2 : Test Statistic and p-value

We need to analyze the result of **One-way ANOVA Test** to make our decision.

Now, let's test the ANOVA and get the summary:

```
aov_advertising <- aov(linear_model)
summary(aov_advertising)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Day         4  146.8   36.71    7.519 6.61e-05 ***
## Residuals   55  268.5    4.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence, the test statistic is **7.519** and p-value is **6.61e-05**.

Step 3 : Statistical Decision

Therefore, we **reject** Null Hypothesis, H_0 because $p\text{-value}$ is less than α , level of significance (that is, $p\text{-value} < 0.01$).

Step 4 : Conclusion

There is enough evidence to support the alternative hypothesis, and therefore we conclude that there is significant difference in the average number of inquiries by day, is accurate.

b) If “yes” to a), which day would you prefer to advertise on if maximizing inquiries was your objective?

To check which day would you prefer to advertise on if maximizing inquiries was your objective, we will use Tukey HSD - Statistical Test for Differences.

```
# Tukey for multiple comparisons of means
TukeyHSD(aov_advertising)

##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = linear_model)
##
## $Day
##              diff          lwr          upr          p adj
## Tuesday-Monday    0.4166667 -2.1273194  2.96065269  0.9903820
## Wednesday-Monday  0.08333333 -2.4606527  2.62731936  0.9999830
## Thursday-Monday   -2.08333333 -4.6273194  0.46065269  0.1573669
## Friday-Monday      2.83333333  0.2893473  5.37731936  0.0218355
## Wednesday-Tuesday -0.33333333 -2.8773194  2.21065269  0.9959087
## Thursday-Tuesday  -2.50000000 -5.0439860  0.04398603  0.0563243
## Friday-Tuesday     2.41666667 -0.1273194  4.96065269  0.0702094
## Thursday-Wednesday -2.16666667 -4.7106527  0.37731936  0.1301276
## Friday-Wednesday   2.75000000  0.2060140  5.29398603  0.0279337
## Friday-Thursday    4.91666667  2.3726806  7.46065269  0.0000118
```

Here, we can observe that there is a significance difference in the mean inquiries of Friday-Thursday group. As the $p - value < 0.01$, the mean inquiries on Friday is greater than mean inquiries on Thursday.

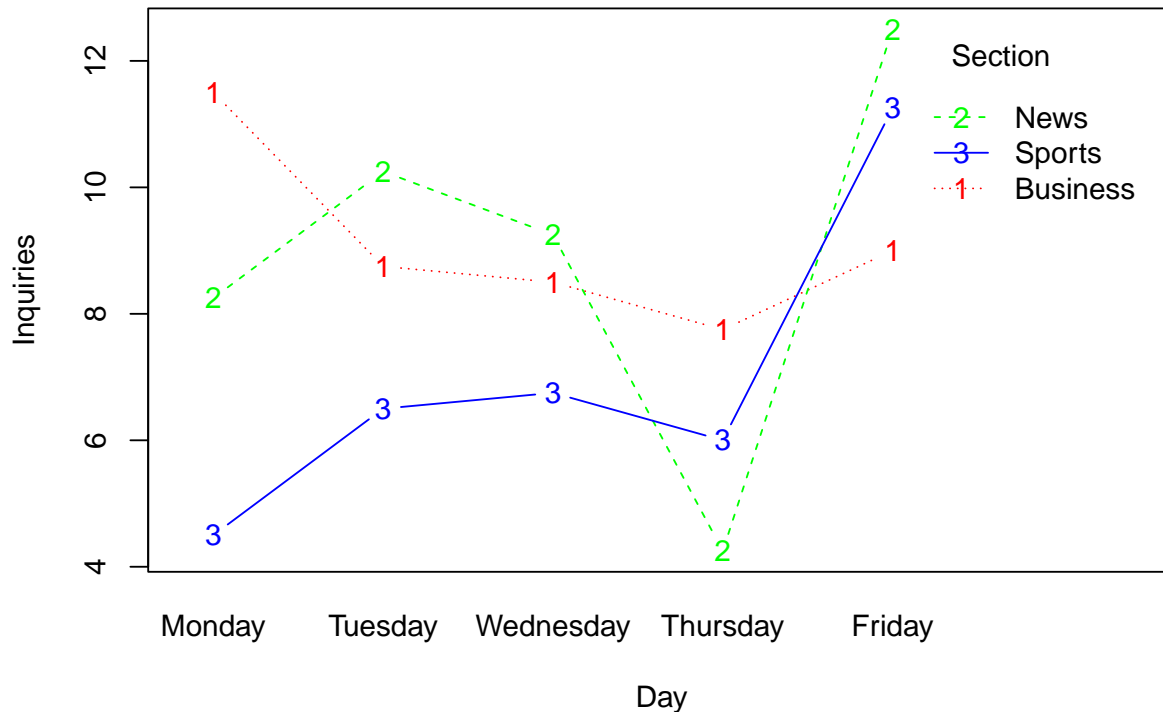
Hence, we can conclude that Friday has a significantly higher mean number of inquiries compared to all other days and we could choosing **Friday** to advertise our ad on.

c) Create an interaction plot for Day and Section on Inquiries. Based on this plot, does there appear to be an interaction; why or why not? (see the examples in the help file for `interaction.plot` on how to create these plots)

Below is an interaction plot for “Day” and “Section” on “Inquiries”.

```
# Interaction Plot
with(advertising_data,
  interaction.plot(x.factor = Day, trace.factor = Section
    , response = Inquiries, type = 'b', ylab = "Inquiries"
    , col = c("red","green","blue")
    , main = "Interaction plot for Day and Section on Inquiries")
)
```

Interaction plot for Day and Section on Inquiries



Yes, Definitely! An interaction between “Day” and “Section” on “Inquiries” can be seen as the lines converge and cross each other and there will be significant increase in inquiries of all sections, particularly “News” section on Fridays.

d) **Test for an interaction between Day and Section. Does your result have an effect on your answer in part a) of this question? Bonus for determining which section and day combination sees the most inquiries using emmeans.**

We will use **Two-Way ANOVA Test** to check whether there is an interaction between the Day and Section.

Step 1 : Hypothesis & Assumptions

Step 1a: Hypothesis The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no significant interaction between Day and Section
vs.

H_A : There is significant interaction between Day and Section

Step 1b: Assumptions

We need to check the assumptions of:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Normality** : The residuals should be nearly normal distributed and show normality. It will be checked by fitting the linear model and using Q-Q Plot and Shapiro-Wilk's Test of model residuals.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals). It will be checked by using Residual vs Fitted plot of model residuals and Levene's Test

In Part (a), we have checked all the assumptions and are satisfied. Hence, we can proceed with the testing using the parametric approach.

Step 2 : Test Statistic and p-value

We need to analyze the result of **Two-Way ANOVA Test** to make our decision.

Now, let's test the ANOVA and get the summary:

```
# Creating the linear model
linear_model_one <- lm(Inquiries ~ Day * Section, data = advertising_data)
# Performing ANOVA
summary(aov(linear_model_one))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Day           4 146.83   36.71   20.910 8.52e-10 ***
## Section       2  53.73   26.87   15.304 8.50e-06 ***
## Day:Section   8 135.77   16.97    9.667 1.12e-07 ***
## Residuals    45  79.00    1.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence, the test statistic is **9.667** and p-value is **1.12e-07**.

Step 3 : Statistical Decision

Therefore, we **reject** Null Hypothesis, H_0 because p-value is less than α , level of significance (that is, p-value < 0.01).

Step 4 : Conclusion

There is enough evidence to support the alternative hypothesis, and therefore we conclude that there is a significant interaction between Day and Section, is accurate.

Reason:

Yes, Definitely! Our result have an effect on our answer in part (a) of this question. Any chances of interaction between “Day” and “Section” has been ignored in part (a). Also, there is an interaction between “Day” and “Section” which affects the result.

EMMEANS

To check which section and day combination sees the most inquiries, we will use *emmeans()* function:

```
library(emmeans)
```

```
## Warning: package 'emmeans' was built under R version 4.2.3
```

```
ans <- emmeans(linear_model_one, pairwise ~ Day:Section)
ans$emmeans
```

```
## Day      Section emmean    SE df lower.CL upper.CL
## Monday   Business 11.50 0.662 45    10.17    12.83
## Tuesday  Business  8.75 0.662 45     7.42    10.08
## Wednesday Business  8.50 0.662 45     7.17     9.83
## Thursday Business  7.75 0.662 45     6.42     9.08
## Friday   Business  9.00 0.662 45     7.67    10.33
## Monday   News      8.25 0.662 45     6.92     9.58
## Tuesday  News     10.25 0.662 45     8.92    11.58
## Wednesday News      9.25 0.662 45     7.92    10.58
## Thursday News      4.25 0.662 45     2.92     5.58
## Friday   News     12.50 0.662 45    11.17    13.83
## Monday   Sports    4.50 0.662 45     3.17     5.83
## Tuesday  Sports    6.50 0.662 45     5.17     7.83
## Wednesday Sports    6.75 0.662 45     5.42     8.08
## Thursday Sports    6.00 0.662 45     4.67     7.33
## Friday   Sports   11.25 0.662 45     9.92    12.58
##
## Confidence level used: 0.95
```

Here, we can see that the pair, that is, Friday and News (“Day” and “Section” respectively) has maximum emmean value of **12.5**. Hence, Friday : News (Day : Section) sees the maximum number of inquiries.

e) The newspaper currently charges the same amount per ad for any day and any section in the newspaper. Assuming you work for the paper, do you have any recommendations for the paper in terms of pricing?

A recommendation for the paper in terms of pricing would include increasing the profitability. To gain more profits we can first observe that on Fridays we see more inquiries than any other day. Therefore to gain profits one might consider increasing the prices of advertisements for Fridays. In addition to this, we can also decrease the prices for news section as we observed less inquiries for this section. Decreasing the prices would definitely attract more people toward this section.

Question 2: A Life of Leisure1 Part I - Model Fitting

It has long been argued that workers with more leisure time are more productive during their work hours. Jay has been trying to convince his employer that he should be allowed to work a four day week for the same annual salary, but so far they have been resistant and require cold hard statistical proof that reduced working hours will help the company's bottom line. After all, Jay is suggesting that he work less while still making the same amount of money. Jay finds data on OECD countries' GDP produced per hour worked which was collected by researchers at a prestigious university (contained in *leisure.csv*) and hires you to analyze the data. This dataset has two columns of interest for us: GDP per hour worked and hours worked.

Firstly, let's loading the data set and check the structure of *leisure.csv*.

```
# Importing the dataset
leisure_data <- read.csv("leisure.csv")
# Let's view the str()
str(leisure_data)

## 'data.frame':  41 obs. of  3 variables:
## $ Country: chr  "AUS" "AUT" "BEL" "CAN" ...
## $ hours  : num  1694 1442 1493 1685 1753 ...
## $ gdp    : num  48.3 15 35.3 41.3 48.3 ...
```

a) Estimate the linear relationship between GDP per hour and hours worked using a linear model. (This includes checking model assumptions using diagnostic plots.) Comment on how well the model fits the data.

To estimate the linear relationship between GDP per hour and hours worked using a linear model, we will conduct *Linear Regression Model* using the two parameters (β_0 is intercept of the model and β_1 is slope of the model):

$$GDP = \beta_0 + \beta_1 * Hours$$

Fitting the Model

Now let's fit the model:

```
lm <- lm(gdp ~ hours, data = leisure_data)
lm

##
## Call:
## lm(formula = gdp ~ hours, data = leisure_data)
##
## Coefficients:
## (Intercept)      hours
##    -5.77909      0.02499
```

Here, the intercept, $\beta_0 = -5.779$ and slope, β_1 is **0.025**.

$$GDP = (-5.779) + (0.025 * Hours)$$

Assumptions

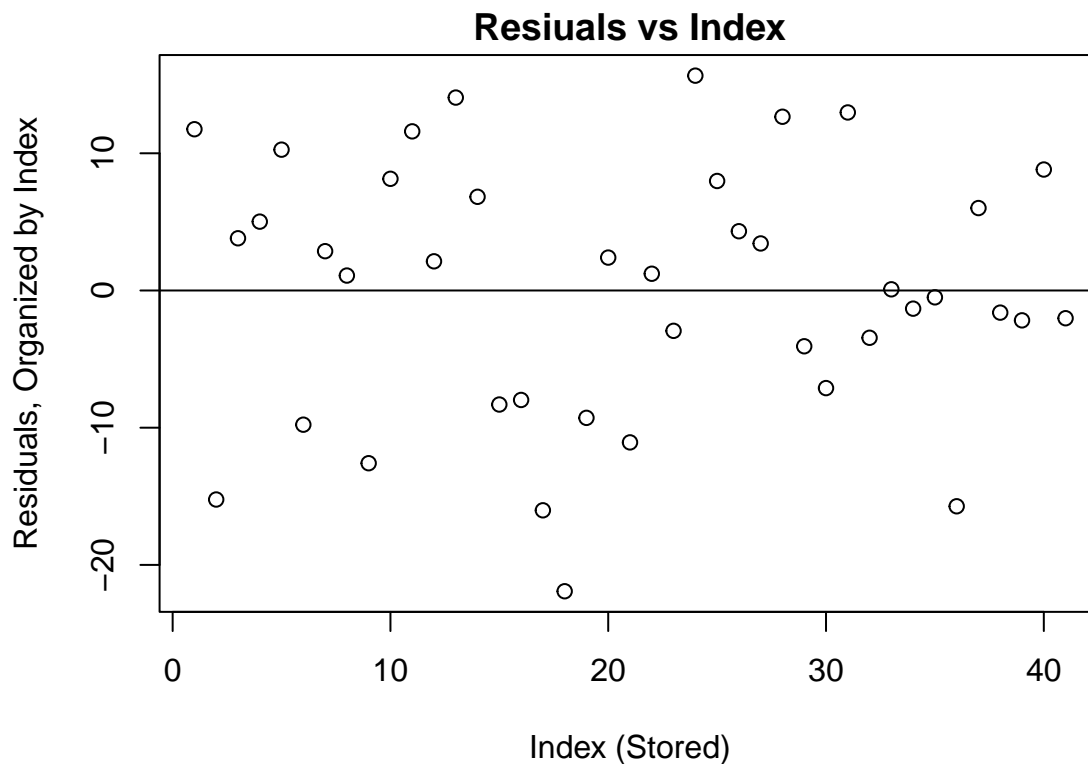
We need to check the assumptions for linear model and tell how well the model fits the data:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Linearity** : There should be a linear relationship between the two variables, that is, “gdp” and “hours”
- **Normality** : The residuals should be nearly normal distributed and show normality.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals).

Now, we will check the assumptions.

Firstly, let's check for *independence*:

```
par(mar = c(4,4,1.5, 1.5), mgp = c(3, 1, 0))
plot(x = 1:length(lm$residuals), y = lm$residuals
, xlab = "Index (Stored)", ylab = "Residuals, Organized by Index", main = "Residuals vs Index")
abline(h = 0)
```



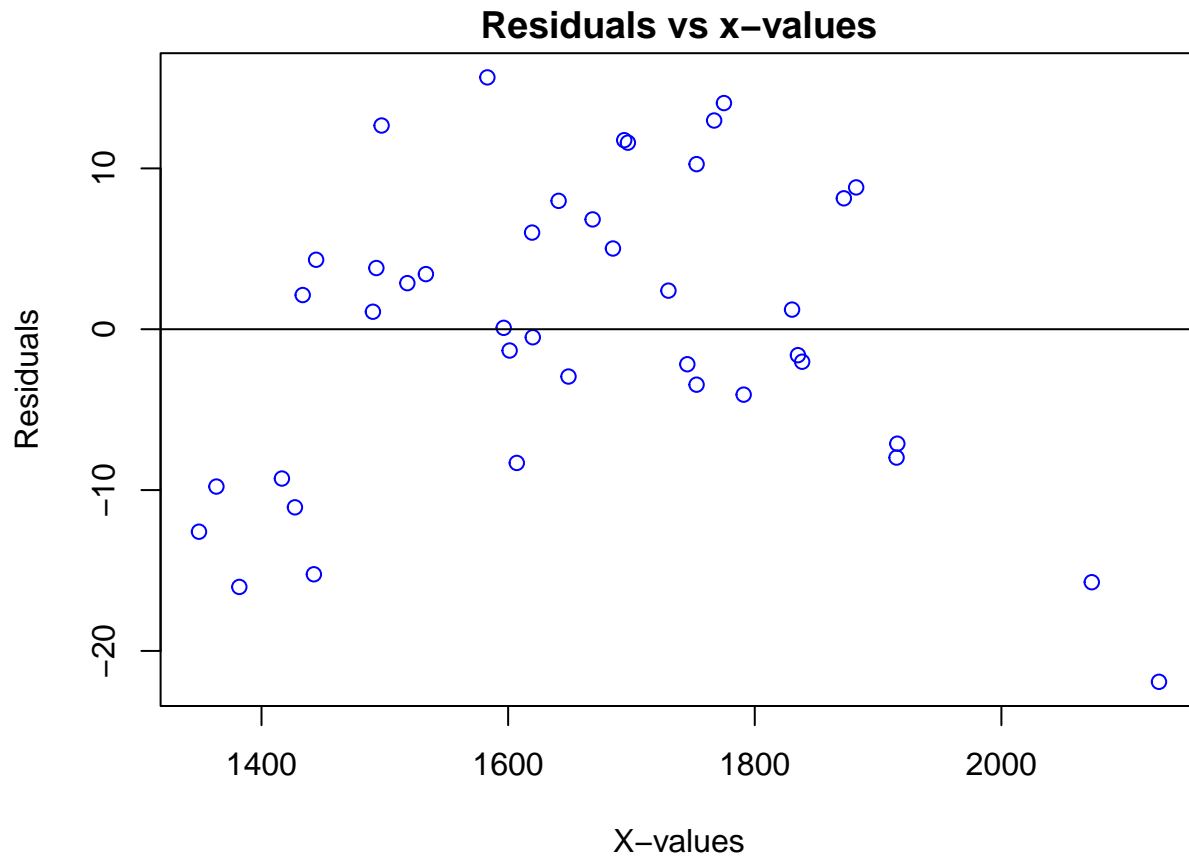
It is **independent**, as no serial correlation can be observed.

Now, we will check the assumptions of *Linearity* among “residuals” and “hours”.

```

par(mar = c(4,4,1.5, 1.5), mgp = c(3, 1, 0))
plot(x = leisure_data$hours, y = lm$residuals
     , xlab = "X-values", ylab = "Residuals", col = "blue"
     , main = "Residuals vs x-values")
abline(h = 0)

```



Here, it can be seen that the residuals vs x are nearly linear. Hence, the assumption is met.

Now, we will check the assumptions of *Normality*:

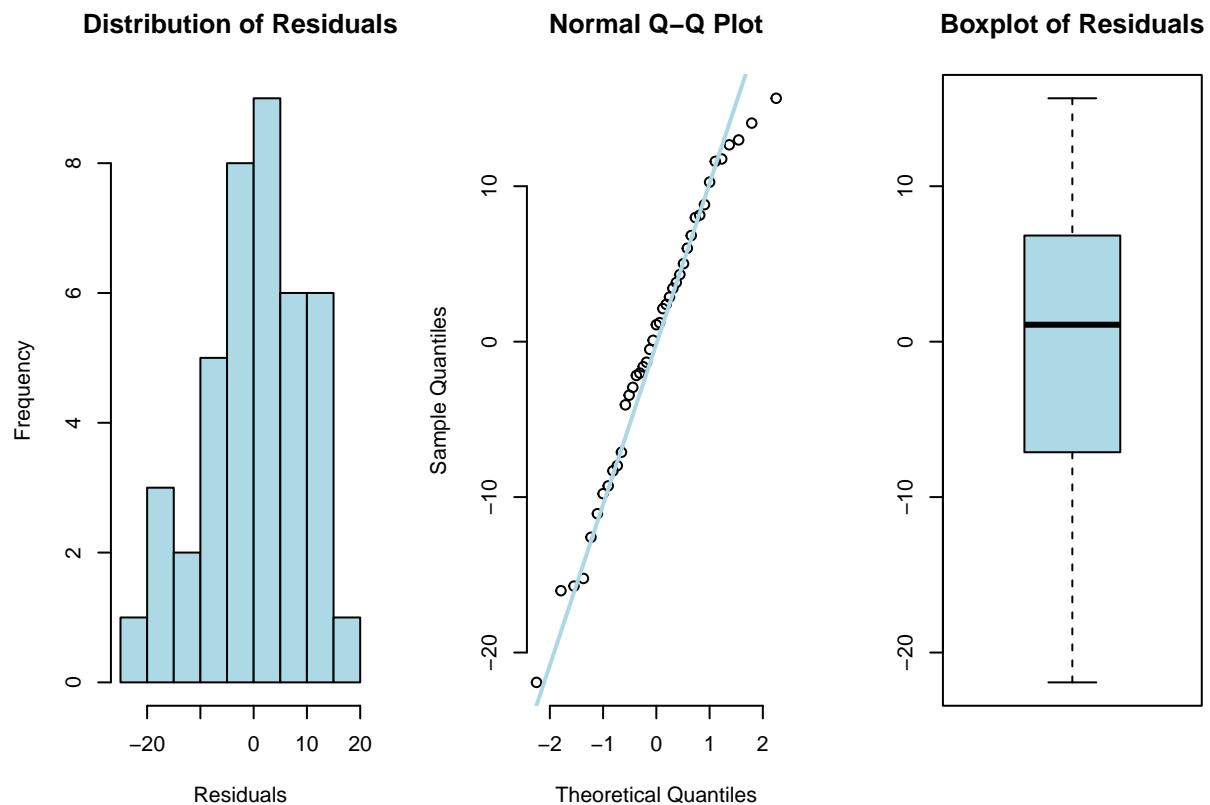
```

par(mfrow=c(1,3))
# Histogram of Residuals
hist(lm$residuals, main = "Distribution of Residuals"
     , col = "lightblue", xlab = "Residuals")

qqnorm(lm$residuals, pch = 1, frame = FALSE)
qqline(lm$residuals, col = "lightblue", lwd = 2)

# Boxplot of Residuals
boxplot(lm$residuals, col = "lightblue"
        , main = "Boxplot of Residuals")

```



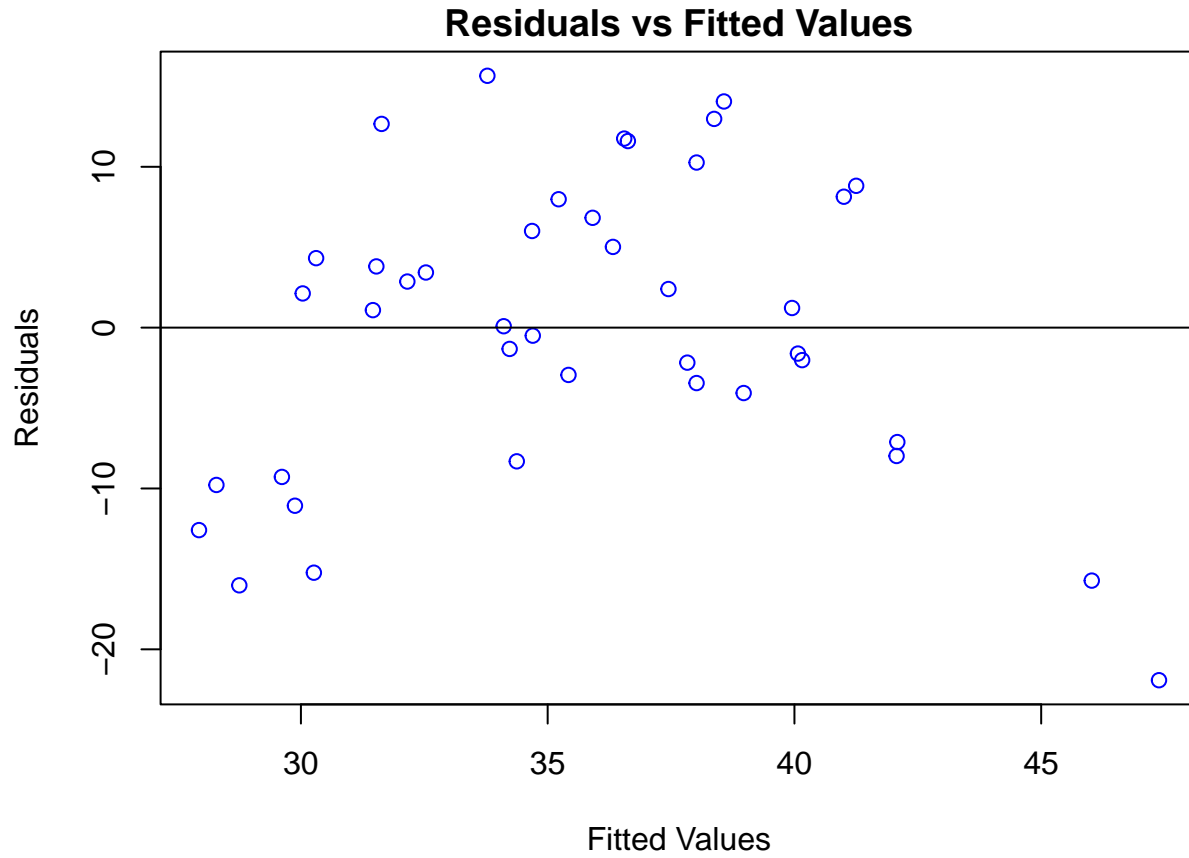
```
# Shapiro-Wilk's normality test
shapiro.test(lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm$residuals
## W = 0.97542, p-value = 0.5083
```

Here, it can be seen that the distribution's frequency peaks at the center and is evenly distributed at the ends, and is confirmed by Shapiro-Wilk's test, as $p\text{-value} > 0.05$. Hence, the assumption of *normality* is met.

Finally, we will check the assumptions of *variability*:

```
par(mar = c(4,4,1.5, 1.5), mgp = c(3, 1, 0))
plot(x = lm$fitted.values, y = lm$residuals, col = "blue"
     , xlab = "Fitted Values", ylab = "Residuals"
     , main = "Residuals vs Fitted Values")
abline(h = 0)
```



The data across the line are nearly constant, as the fitted values increase the residuals are constant in “Residuals vs Fitted Values”. Hence, the condition of *variability* is met.

Hence, **all of the conditions are met** and we can assume **the model to be a good fit** for the data.

b) Conduct a hypothesis test on the slope parameter of the model in a) to determine if there is a linear relationship between GDP per hour and hours worked.

Step 1 : Hypothesis & Assumptions

Step 1a: Hypothesis The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no significant linear relationship between GDP per hour and hours worked.

vs.

H_A : There is significant linear relationship between GDP per hour and hours worked.

Here, we will consider slope Parameter equal to zero for H_0 , Null hypothesis and for H_A , Alternative hypothesis slope Parameter is not zero.

Step 1b: Assumptions

We need to check the assumptions of:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Linearity** : There should be a linear relationship between the two variables, that is, “gdp” and “hours”

- **Normality** : The residuals should be nearly normal distributed and show normality.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals).

All the assumptions are depicted using diagnostic plots in part-a of this question and all assumptions are satisfied. Thus, we will proceed with the further testing to check linearity using the slope parameter of the model:

In Part (a) of the question, we have check all of the assumptions using diagnostic plots (and statistical testing). There we got satisfied result. Hence, we can conduct a hypothesis test on the slope parameter of the model

Step 2 : Test Statistic and p-value

```
summary(lm)

##
## Call:
## lm(formula = gdp ~ hours, data = leisure_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.920  -7.114   1.089   6.828  15.658
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.779086  13.082473  -0.442   0.6611
## hours         0.024986   0.007832   3.190   0.0028 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.398 on 39 degrees of freedom
## Multiple R-squared:  0.207, Adjusted R-squared:  0.1866
## F-statistic: 10.18 on 1 and 39 DF, p-value: 0.002803
```

Here, The p-value of slope parameter is **0.0028**. Hence, the test statistic is **3.190** and p-value is **0.025**.

Step 3 : Statistical Decision

Here, the slope parameter is not equal to zero.

Therefore, we **reject** Null Hypothesis, H_0 because p-value is less than α , level of significance (that is, p-value < 0.05).

Step 4 : Conclusion

There is enough evidence to support the alternative hypothesis, and therefore we conclude that there exists a significant Linear Relationship between GPD per hour and Hours worked, is accurate.

c) You will have hopefully determined that the relationship here is likely not only linear and therefore the model in a) is not the best model for this data. Use a linear model to fit the quadratic relationship between GDP per hour and hours worked. (Again, use diagnostic plots and comment on how well the model fits the data).

Fitting the model

Quadratic Linear Model is used to fit the quadratic relationship between GDP per hour and hours worked.

```
# Quadratic Linear Model
quad_model <- lm(gdp ~ hours + I(hours^2), data = leisure_data)
quad_model
```

```
##
## Call:
## lm(formula = gdp ~ hours + I(hours^2), data = leisure_data)
##
## Coefficients:
## (Intercept)      hours      I(hours^2)
## -4.349e+02    5.385e-01   -1.517e-04
```

Here, after looking the intercept and coefficients variable, the quadratic model will be:

$$GDP = (-434.90) + 0.538 * Hours + (-0.00015) * Hours^2$$

Assumptions

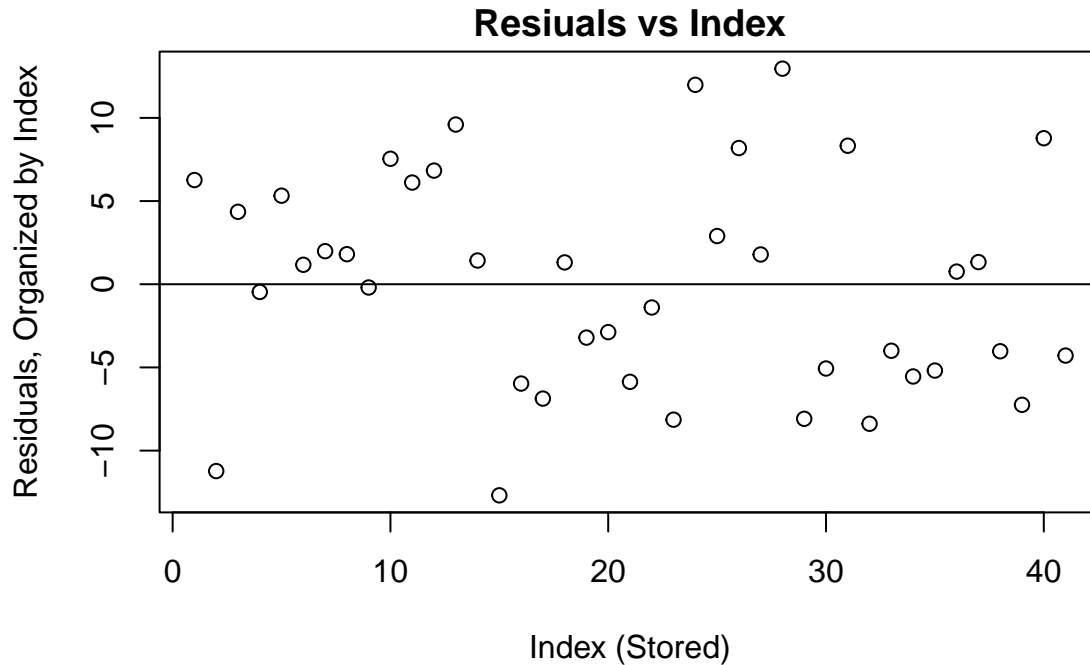
We need to check the assumptions for quadratic linear model and tell how well the model fits the data:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Linearity** : There should be a linear relationship between the two variables, that is, “gdp” and “hours”
- **Normality** : The residuals should be nearly normal distributed and show normality.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals).

Now, we will check the assumptions.

Firstly, let's check for *independence*:

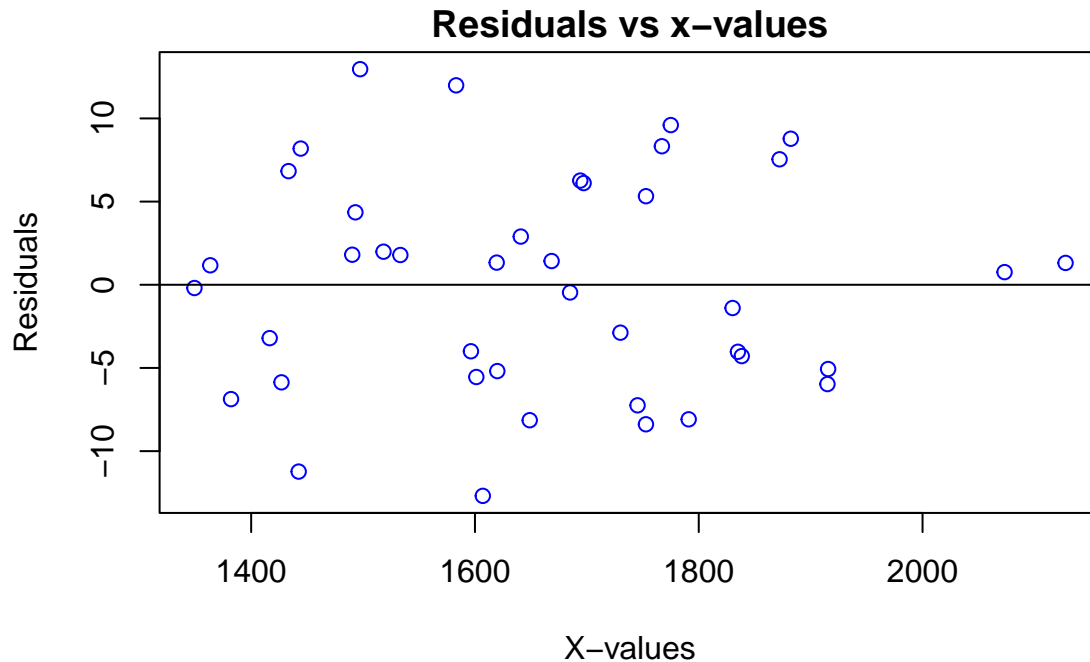
```
par(mar = c(4,4,1.5, 1.5), mgp = c(3, 1, 0))
plot(x = 1:length(quad_model$residuals), y = quad_model$residuals
     , xlab = "Index (Stored)"
     , ylab = "Residuals, Organized by Index"
     , main = "Residuals vs Index")
abline(h = 0)
```



It is **independent**, as no serial correlation can be observed.

Now, we will check the assumptions of *Linearity* among “hours” and “residuals”.

```
par(mar = c(4,4,1.5, 1.5), mgp = c(3, 1, 0))
plot(x = leisure_data$hours, y = quad_model$residuals
     , xlab = "X-values", ylab = "Residuals", col = "blue"
     , main = "Residuals vs x-values"); abline(h = 0)
```



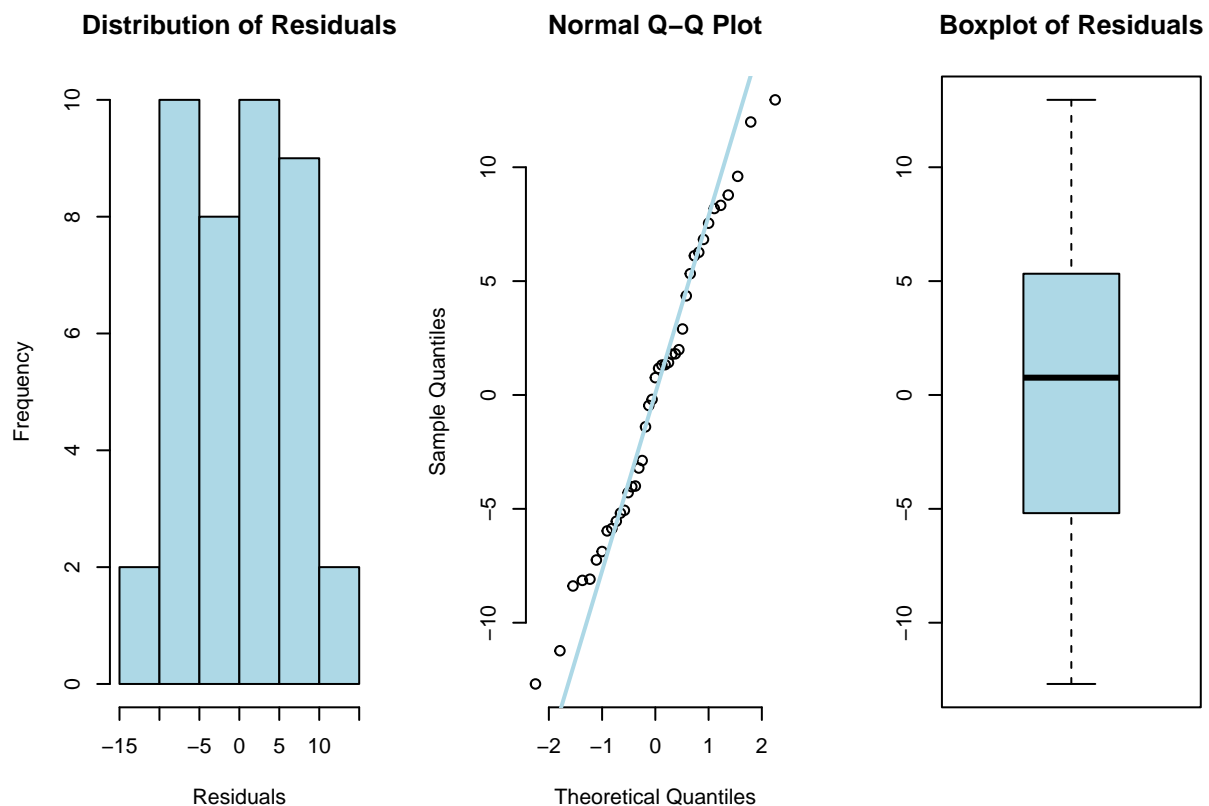
Here, it can be seen that the residuals vs x-value are nearly linear, as no significant pattern for residuals can be observed. Hence, the assumption is met.

Now, we will check the assumptions of *Normality*:

```
par(mfrow=c(1,3))
# Histogram of Residuals
hist(quad_model$residuals, main = "Distribution of Residuals"
     , col = "lightblue", xlab = "Residuals")

qqnorm(quad_model$residuals, pch = 1, frame = FALSE)
qqline(quad_model$residuals, col = "lightblue", lwd = 2)

# Boxplot of Residuals
boxplot(quad_model$residuals, col = "lightblue"
       , main = "Boxplot of Residuals")
```



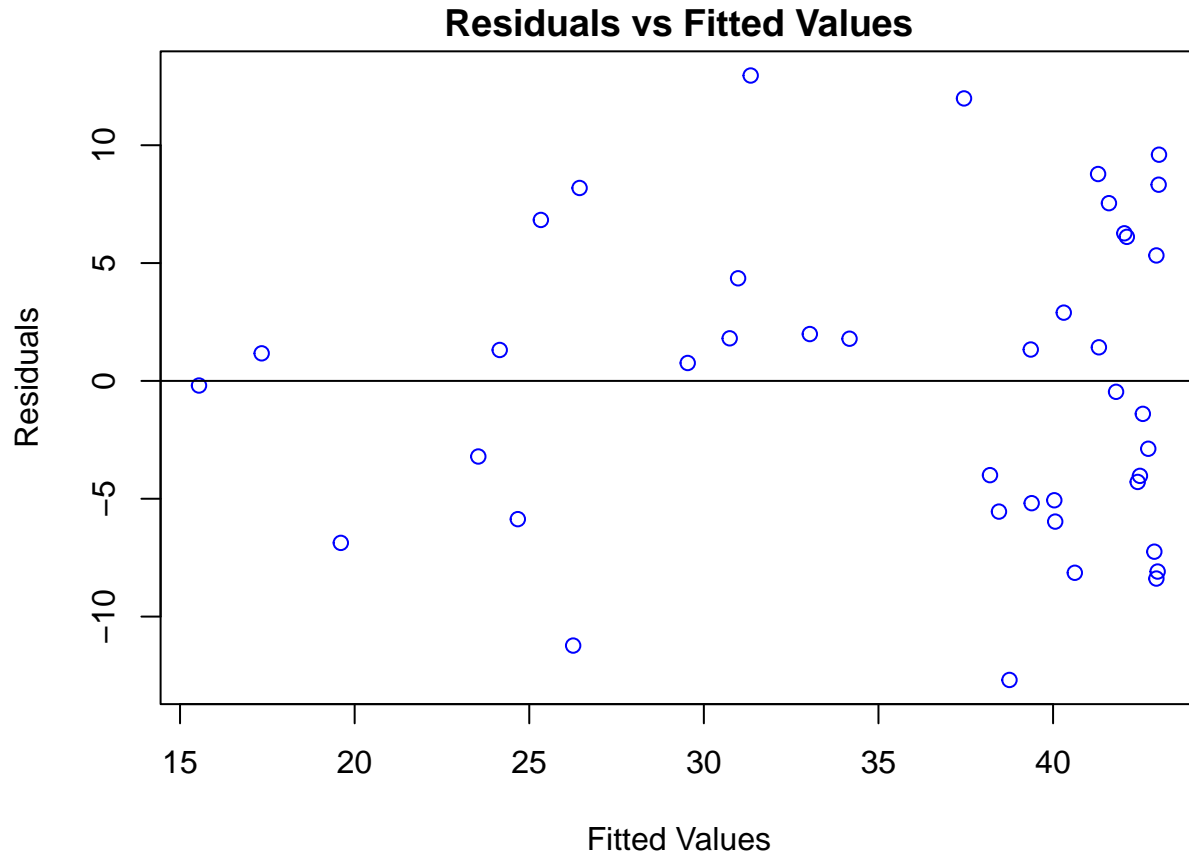
```
# Shapiro-Wilk's normality test
shapiro.test(quad_model$residuals)
```

```
##
## Shapiro-Wilk normality test
##
## data:  quad_model$residuals
## W = 0.97625, p-value = 0.5371
```

Here, it can be seen that the distribution's the frequency peaks at the center and is evenly distributed at the ends, and is confirm by Shapiro-Wilk's test, as $p\text{-value} > 0.05$. Hence, the assumption of *normality* is met.

Finally, we will check the assumptions of *variability*:

```
par(mar = c(4,4,1.5, 1.5), mgp = c(3, 1, 0))
plot(x = quad_model$fitted.values, y = quad_model$residuals, col = "blue"
     , xlab = "Fitted Values", ylab = "Residuals"
     , main = "Residuals vs Fitted Values")
abline(h = 0)
```



The data across the line are nearly constant, as the fitted values increase the residuals are constant in “Residuals vs Fitted Values”. Hence, the condition of *variability* is met.

Hence, **all of the conditions are met** and we can assume **the model to be a good fit** for the data.

Now, let's summary the quadratic regression model

```
summary(quad_model)
```

```
##
## Call:
## lm(formula = gdp ~ hours + I(hours^2), data = leisure_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6889  -5.1879   0.7604   5.3237  12.9585
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.349e+02  6.872e+01  -6.328 2.02e-07 ***
## hours       5.385e-01  8.168e-02   6.593 8.79e-08 ***
## I(hours^2)  -1.517e-04  2.407e-05  -6.301 2.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.658 on 38 degrees of freedom
## Multiple R-squared:  0.6122, Adjusted R-squared:  0.5918
## F-statistic: 29.99 on 2 and 38 DF,  p-value: 1.526e-08
```

After looking the assumptions (using diagnostic plots) and the summary statistics of the quadratic regression model, we can say that the model is linear as well as quadratic (linear is done in above part of this question). The intercept, hours and quadratic term ($hours^2$) is **-4.349e+02**, **5.385e-01** and **-1.517e-04** respectively. Hence, the negative value of quadratic term ($hours^2$) explain about the effect of hours worked on GDP per hour diminishes as hours worked increase.

d) Plot the data and your model from a) and c) all on the same plot. Include a legend.

Given is the plot of Linear and Quadratic Model:

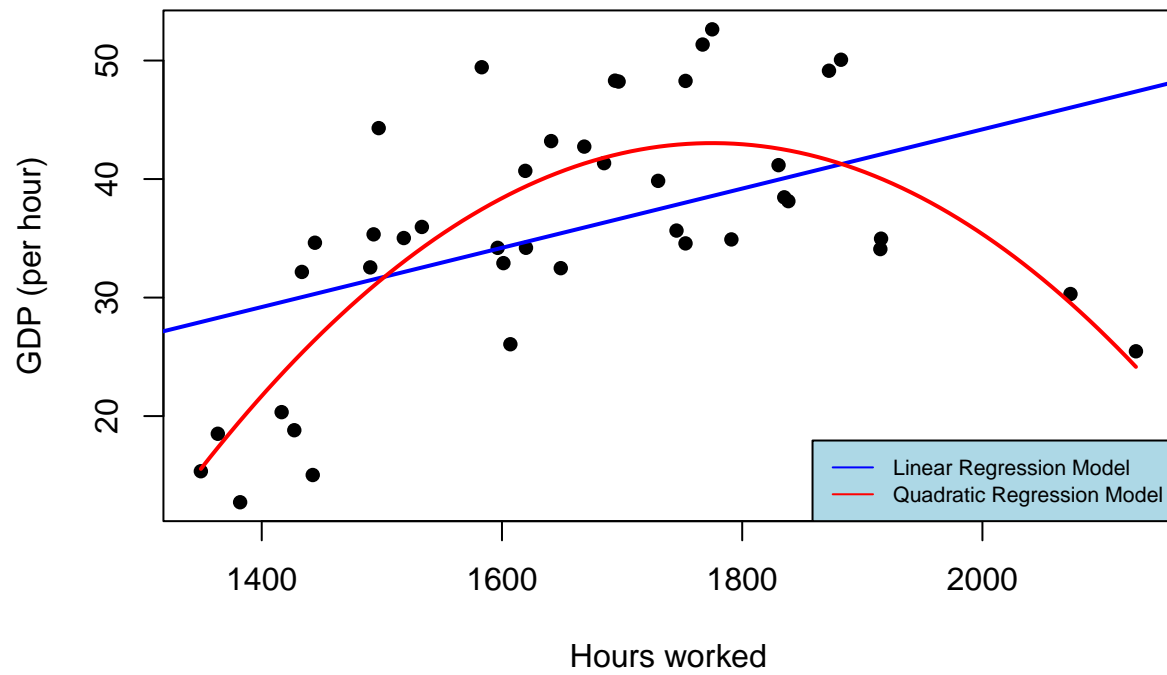
```
# Plotting the data points
# from leisure.csv
plot(gdp ~ hours, data = leisure_data
     , main = "Model Comparison between GDP (per hours) and Hours Worked"
     , xlab = "Hours worked"
     , ylab = "GDP (per hour)", pch=16)

# Linear Model's Regression Line
abline(lm, col = "blue", lw=2)

# Quadratic Model curve
curve(predict(quad_model
             , newdata = data.frame(hours = x, hours2 = x^2))
      , add = TRUE, col = "red", lw=2)

# Legend
legend(x = "bottomright"
      , legend=c("Linear Regression Model", "Quadratic Regression Model")
      , col=c("blue", "red"), cex=0.7
      , lty=1:1, bg='lightblue')
```

Model Comparison between GDP (per hours) and Hours Worked



Question 3: A Life of Leisure Part II - Interpretation

For this question “The Model” is your model from Question 2c).

a) Explain why The Model in Q2c) is better than your model in Q2a). There are MANY comparisons to make; to get full marks you must make at least 4.

The model in question-2 (c), that is, *Quadratic Regression Model* is better than our model in question-2 (a), that is, *Linear Regression Model*. Below are the comparisons:

1. **Smaller p -value:** As we know, low p -value (typically less than 0.05) indicates that the coefficient is statistically significant and contributes significantly to the model's predictive power. In the case of quadratic regression, the p -value for the quadratic term should be low, indicating that it adds significant value to the model compared to a linear model. In our case, the p -value of quadratic regression model is **1.526e-08** and the p -value of linear regression model is **0.002803**. Hence, quadratic regression model has smaller p -value than the linear regression model which depicts a significant model fit.
2. **Higher Adjusted R^2 :** In general, a quadratic model should have a higher adjusted R^2 compared to a linear model, as it can capture more complex relationships between the independent and dependent variables. In question 2, we got adjusted R-squared as **0.5918** in quadratic regression model and **0.1866** in linear regression model. Hence, quadratic regression model has higher Adjusted R^2 value than the linear regression model which depicts a better model fit and higher predictive power.
3. **Lower Residual Standard Error (RSE):** In general, a quadratic model should have a lower RSE compared to a linear model, as it can capture more of the variation in the data. As in the above question, we got residual standard error(RSE) as 6.658 and 9.398 in quadratic regression model and linear regression model respectively. Thus, a quadratic model has lower RSE compared to a linear model which depicts that quadratic model has more precise prediction intervals and has more variation in the data.
4. **Parsimonious Model:** The quadratic regression model is parsimonious, because it includes only the necessary predictor variables. Moreover, it has all the coefficient terms, high adjusted R^2 value, smaller p -value, and lower Residual Standard Error (RSE). On the other hand, linear regression model has more p -value, high Residual Standard Error (RSE), and low adjusted R^2 value than quadratic regression model. Hence, quadratic regression model is a better fit model.

b) Give The Model equation.

The quadratic model equation is given by:

$$GDP = (-434.90) + 0.538 * Hours + (-0.00015) * Hours^2$$

c) Based on The Model equation in b), what is the average number of work hours per person that would maximize GDP per hour?

To get the average number of work hours per person that would maximize GDP per hour, we would use the *vertex of parabola* equation, which is given by:

$$vertex = -b/2a$$

Here,

a = coefficient of quadratic term ($hours^2$) in quadratic model(i.e, **-0.00015**), and

b = coefficient of linear term (*hours*) in quadratic model (i.e, **0.538**)

Let's insert the value in the formula and we get:

```
vertex_of_parabola <- (-0.538) / (2*(-0.00015))  
print(vertex_of_parabola)
```

```
## [1] 1793.333
```

Hence, the average number of work hours per person is **1793.33** that would maximize GDP per hour.

d) Jay currently works 1992 hours a year. As mentioned, he wants to ask his employer for a four-day work week. Would your analysis convince an employer to move their employees to a four-day work week? Explain your process. If not, is there still a way for Jay to work less than he currently is and still make the same amount of money? What type of compromise could be offered?

Through our analysis, we concluded that a person would need to work for an average of 1793.33 hours to maximize the GDP per hour. We have been given that Jay is working currently for 1992 hours which is clearly more than the average number of hours mentioned above. But according to our analysis this would result in decreasing the productivity only. In other words we can say that in order to maximize the GDP per hour, an employee would require to work for 1793.33 hours annually or 34.4 hours weekly. This would mean that in a four day week he would have to work 8.59 hours each day to maintain the optimum level of GDP per hour.