COLLEGE OF COMPUTIG

DEPARTMENT OF DATA SCIENCE

REGRATION PROGECT

GROUPMEMBERS NAME	**ID**
1. NATNAEL BEKELE	DBU1501407
2. HAFIZ HUSSEN	DBU1501241
3. REDIET ESUBALEW	DBU1501704
4. BEREKET GETAW	DBU1501044
5. DAWIT ALEMU	DBU1501117
6. YIFERU MEKONEN	DBU1501562
7. ELBETEL ABEDI	DBU1501145
8. GENET MINDA	DBU1501217

1, Use mtcars data from R and fit a multiple linear regression miles per gallon (fuel efficiency) on horsepower, weight (in 1000 lbs), 1/4 mile time, number of cylinders and displacement (mpg on hp , wt , qsec, cyl and disp).

---Here is our code---

```
# Load necessary libraries
library(lmtest) #For Breusch-Pagan test
library(car)# For Variance Inflation Factor (VIF)
library(nlme) # For Weighted Least Squares (WLS)

#load the mtcars data set
    head(mtcars)
```

output:

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
    dim(mtcars)
```

output:

```
> dim(mtcars)
[1] 32 11
```

Explanation:

The output dim(mtcars) in R indicates that the mtcars dataset has 32 rows and 11 columns. This means it contains data on 32 observations (e.g., different car models) across 11 variables (e.g., mpg, horsepower, weight). The dim() function is used to display the structure of datasets, where the first number represents the rows and the second represents the columns.

fitting the model####

fit=lm(mpg~hp+wt+qsec+cyl+disp,mtcars)

summary(fit)

```
Call:
lm(formula = mpg ~ hp + wt + qsec + cyl + disp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3117 -1.3483 -0.4352  1.2603  5.6094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.87361    9.91809   3.617  0.00126 **
hp          -0.01584    0.01527  -1.037  0.30908
wt          -4.22527    1.25239  -3.374  0.00233 **
qsec         0.25382    0.48746   0.521  0.60699
cyl         -1.15608    0.71525  -1.616  0.11809
disp         0.01195    0.01191   1.004  0.32484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.547 on 26 degrees of freedom
Multiple R-squared:  0.8502,    Adjusted R-squared:  0.8214
F-statistic: 29.51 on 5 and 26 DF,  p-value: 6.182e-10
```

Explanation:

Cal
l

- This shows the formula used in the regression model:
- mpg ~ hp + wt + qsec + cyl + disp
    - The dependent variable is mpg (miles per gallon).
    - The independent variables are hp (horsepower), wt (weight in 1000 lbs), qsec (1/4 mile time), cyl (number of cylinders), and disp (engine displacement).

Residuals

Min 1Q Median 3Q Max

-4.3117 -1.3483 -0.4352 1.2603 5.6094

- Residuals are the differences between the observed and predicted values of mpg.
- The values tell us the range of residuals:
- The smallest residual (Min): -4.3117 (the model underestimates the mpg by this much).
- The largest residual (Max): 5.6094 (the model overestimates the mpg by this much).
- The Median is -0.4352, meaning that the model tends to slightly underestimate mpg on average.
- Ideally, the residuals should be symmetrically distributed around zero.

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.87361	9.91809	3.617	0.00126
hp	-0.01584	0.01527	-1.037	0.30908
wt	-4.22527	1.25239	-3.374	0.00233
qsec	0.46285	0.48746	0.521	0.60699
cyl	-1.15608	0.71525	-1.616	0.11809
disp	0.01195	0.01191	1.004	0.32484

Each row explains the effect of a predictor variable on mpg:

- **(Intercept):**
  - The estimated intercept is 35.87361. This is the predicted mpg when all predictors are zero.
  - This value has little practical meaning because predictors like weight (wt) and horsepower (hp) cannot be zero.
- **hp (Horsepower):**
  - Coefficient: -0.01584
    - For every 1-unit increase in horsepower, mpg decreases by approximately 0.01584 units, holding all other variables constant.
  - p-value: 0.30908
    - This is greater than 0.05, meaning hp is not a significant predictor of mpg in this model.
- **wt (Weight):**
  - Coefficient: -4.22527
    - For every 1000-lb increase in vehicle weight, mpg decreases by 4.22527 units, holding all other variables constant.
  - p-value: 0.00233
    - This is less than 0.05, so weight is a significant predictor of mpg.
- **qsec (1/4 Mile Time):**
  - Coefficient: 0.46285
    - For every 1-se cond increase in qsec, mpg increases by 0.46285 units, holding other variables constant.
  - p-value: 0.60699
    - This is greater than 0.05, so qsec is not a significant predictor of mpg.
- **cyl (Cylinders):**
  - Coefficient: -1.15608
    - For every additional cylinder, mpg decreases by 1.15608 units, holding other variables constant.

    - p-value: 0.11809
        - This is greater than 0.05, so cylinders are not a significant predictor of mpg.

- **disp (Displacement):**
    - Coefficient: 0.01195
        - For every 1-unit increase in displacement, mpg increases by 0.01195 units, holding other variables constant.
    - p-value: 0.32484
        - This is greater than 0.05, so displacement is not a significant predictor of mpg.

Significant codes

The Signif. codes section in the output of an R linear model (lm) summary explains the statistical significance of the predictors in the model, based on their p-values. The codes are as follows:

- ***: Highly significant (p-value < 0.001)
- **: Very significant (p-value < 0.01)
- *: Significant (p-value < 0.05)
- . : Marginally significant (p-value < 0.1)
- Blank: Not significant (p-value >= 0.1)

These codes provide a quick indication of how strongly each predictor variable (e.g., hp, wt, qsec, cyl, disp) influences the dependent variable (mpg in this case).

In the output:

- (Intercept) and wt have **, meaning they are statistically very significant predictors (p-value < 0.01).
- hp, qsec, cyl, and disp do not have any significance codes, indicating that their effects on mpg are not statistically significant at conventional thresholds.

Residual Standard Error (RSE)

Residual standard error: 2.547 on 26 degrees of freedom

- The RSE is 2.547, which measures the average deviation of observed mpg values from the predicted values.
- Smaller RSE values indicate better model accuracy.

R-Squared and Adjusted R-Squared

Multiple R-squared: 0.8502, Adjusted R-squared: 0.8214

R-squared: 0.8502

- About 85.02% of the variability in mpg is explained by the model.

Adjusted R-squared: 0.8214

- Adjusted R-squared accounts for the number of predictors in the model.

F-Statistic

F-statistic: 29.51 on 5 and 26 DF, p-value: 6.182e-10

- F-statistic: 29.51 tests whether the model as a whole is may or may not significant.
- p-value: 6.182e-10 (very small) indicates the model is statistically significant overall. At least one predictor is significantly associated with mpg.

So to know the signification of the model we must check the assumptions.

2, Check the assumptions of the linear regression, including linearity, homoscedasticity, and normality of residuals.

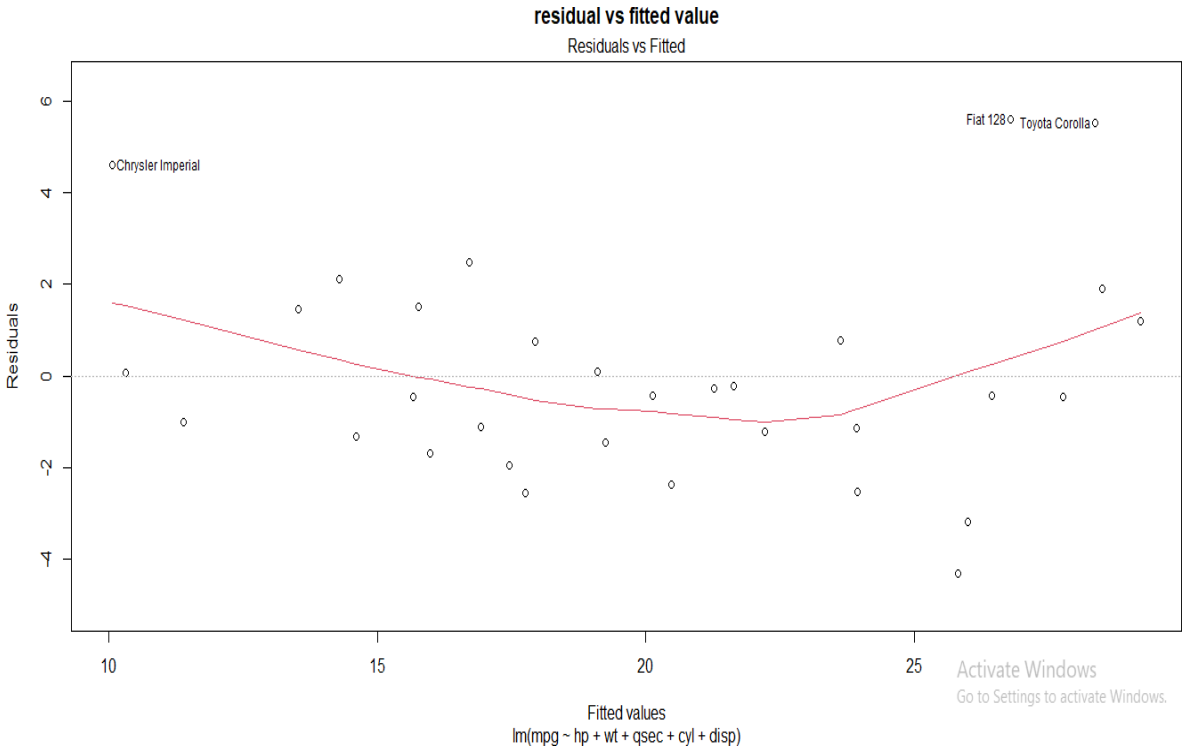
First we can check the assumption of linearity

To check the **linearity assumption** in a multiple linear regression model, we need to ensure that the relationship between the predictors and the response variable (mpg) is linear. This can be done by examining **residual plots** (residuals vs. fitted values).

---Here is our code---

```
###check the assumption of linearity##
# by using residual vs fitted value plot #
plot(fit,which=1,main="residual vs fitted value")
```

output:



Interpretation of the Residuals vs. Fitted Plot:

Axes:

- **X-axis (Fitted values):** The predicted values of `mpg` from the regression model.
- **Y-axis (Residuals):** The differences between observed and predicted values of `mpg`.

• **Non-Linearity Detected:** The red smooth line is curved rather than flat, indicating that the relationship between the predictors and `mpg` is not purely linear. This suggests that a transformation (e.g., polynomial terms or interactions) may be needed.

Second check the assumption of homoscedasticity

For a small sample size like 32 the Goldfeld-Quandt Test is the most appropriate to test the assumption of homoscedasticity. This test is specifically designed for smaller datasets and works by dividing the data into groups and comparing the variances of the residuals.

The Breusch-Pagan Test and White's Test are more suitable for larger datasets, as they rely on asymptotic properties and may not perform well with small sample sizes.

Thus, for our model with 32 observations, the Goldfeld-Quandt Test is the best choice.

---Here is our code---

```
######check constant variance(homosdacity)assumption######
 ### by using goldfeld-quandt test###

gqtest(fit, order.by = ~ hp, data = mtcars, fraction =0.2)
```

out put:

```
        Goldfeld-Quandt test

data:  fit
GQ = 0.75306, df1 = 7, df2 = 6, p-value = 0.6436
alternative hypothesis: variance increases from segment 1 to 2
```

Interpretation of the Goldfeld-Quandt Test:

1. **Test Purpose:**
   - The Goldfeld-Quandt test checks for **heteroscedasticity** by splitting the dataset into two segments and comparing their variance.
2. **Null Hypothesis (H0)**:
   - The variance of the two group are equal
3. **Alternative Hypothesis (Ha)**:
   - The variance of the two group are different (heteroscedasticity is present).

4. **Results:**
    - **GQ Statistic = 0.75306**
    - **Degrees of Freedom:** df1 = 7, df2 = 6
    - **p-value = 0.6436**
5. **Conclusion:**
    - Since **p-value > 0.05**, we **fail to reject H0**.
    - This means **there is no significant evidence of heteroscedasticity**, confirming that variance is likely constant.

Then check the assumption of normality

The Shapiro-Wilk test is often considered the most appropriate because it has high power and is effective for detecting deviations from normality, particularly for small to moderate sample sizes. However, for larger sample sizes, the graphical methods or the Kolmogorov-Smirnov Test can provide useful insights too.

So, In our data set the appropriate methods to test the assumption of normality is shapiro-Wilk test.

---Here is our code---

```
###check assumption of normality###
  ##by using  shapiro wilk test###
    shapiro.test(residuals(fit))
```

output:

```
        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.94137, p-value = 0.08186
```

Interpretation of the Goldfeld-Quandt Test:

1. **Test Purpose:**
    - The **Shapiro-Wilk normality test** is used to check if the residuals of a regression model are normally distributed.
2. **Null Hypothesis (H0)**:
    - The residuals are normally distributed.
3. **Alternative Hypothesis (Ha)**:
    - The residuals are not normally distributed

4. **Results:**

    - **Test Statistic (W)**: 0.94137
    - **p-value**: 0.08186

5. **Conclusion:**

    - Since the **p-value** > **0.05**, we **fail to reject the null hypothesis**.
    - This means there is **no significant evidence** to conclude that the residuals deviate from normality.

3. Check the assumption of multicollinearity

To check **multicollinearity**, various methods exist, but **Variance Inflation Factor (VIF)** is the most commonly used because it directly quantifies the severity of multicollinearity. But other methods like **Correlation matrix** is a **quick check**, but it only detects pairwise correlations.

So, the appropriate methods to check the assumption of multicollinearity is to calculate the variance inflation factor(VIF).

---Here is our code---

check the assumption of multicolliniarity###

by using variance inflation factor#####

vif(fit)

output:

```
       hp        wt      qsec       cyl      disp
 5.235473  7.174933  3.625423  7.796396 10.406548
```

Interpretation:

- **disp (10.41)** has the highest VIF, indicating **high multicollinearity** (VIF > 10 is concerning).
- **cyl (7.79) and wt (7.17)** show moderate multicollinearity (VIF between 5-10).
- **hp (5.23)** has moderate collinearity but is closer to an acceptable range.
- **qsec (3.63)** shows low collinearity (VIF < 5 is generally acceptable).

Based on the VIF values, **multicollinearity is present**, particularly for **disp (10.41)** and moderately for **cyl (7.79) and wt (7.17)**.However, the assumption of **no multicollinearity does not completely fail**—it indicates **some concern** that may need to be addressed.

4. If problems are identified during the residual analysis or VIF check, apply the appropriate remedial measures.

Methods Used to Address Linearity & Multicollinearity

1, Multicollinearity Fixes (VIF Reduction)

❖ **Method Used:**

- **Centering (Mean-Subtraction)**:

```
mtcars$hp_centered <- scale(mtcars$hp, center = TRUE, scale = FALSE)
mtcars$wt_centered <- scale(mtcars$wt, center = TRUE, scale = FALSE)
```

  - Reduces correlation between predictors and improves interpretability.
- **Dropping Highly Correlated Variables (`cyl` & `disp`)**:

```
model_reduced <- lm(mpg ~ hp_centered + wt_centered + qsec, data =
mtcars)
```

  - If `cyl` and `disp` had high VIF values (>10), removing them helps eliminate multicollinearity.
  - If two variables have **correlation > 0.8**, drop one.
- **Variance Inflation Factor (VIF) Check**

```
vif(model_reduced)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
            hp_centered             wt_centered           qsec_centered
               4.924442                2.558084                2.873893
hp_centered:wt_centered
               1.019414
```

  - Ensures all remaining variables have **VIF < 5** (acceptable range).

2, Linearity Fixes (Residual Plot Improvement)

❖ **Method Used:**

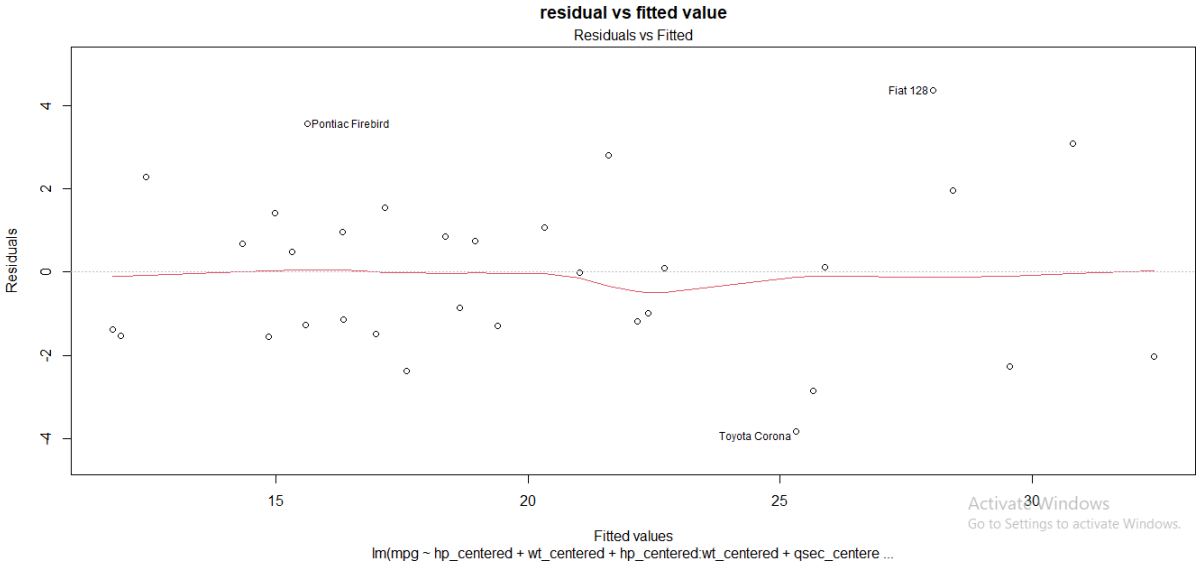
- **Adding Interaction Term (`hp_centered:wt_centered`)**

```
model_reduced <- lm(mpg ~ hp_centered + wt_centered +
hp_centered:wt_centered + qsec, data = mtcars)
```

  - Helps capture possible non-linear effects between `hp` and `wt`.
- **Residuals vs Fitted Plot (Visual Linearity Check)**

```
plot(model_reduced, which = 1)
```

output:



- Ensures residuals are randomly scattered without patterns

☐ Summary of Methods Used

Assumption	Method Used in Code	Why It Helps?
Multicollinearity	Centering (`scale()`)	Reduces correlation
	Dropping highly correlated variables	Removes redundancy
	Checking VIF (`vif()`)	Ensures no high collinearity
Linearity	Adding interaction terms (`hp:wt`)	Captures non-linear relationships
	Checking residuals (`plot()`)	Verifies linearity visually

5, After making any necessary adjustments, refit the model and check the results and make an interpretation about it and draw a conclusion.

---Here is our code---

###Refit the model###

model_reduced <- lm(mpg ~ hp_centered + wt_centered + hp_centered:wt_centered + qsec, data = mtcars)

Recheck the assumption Of Linearitty

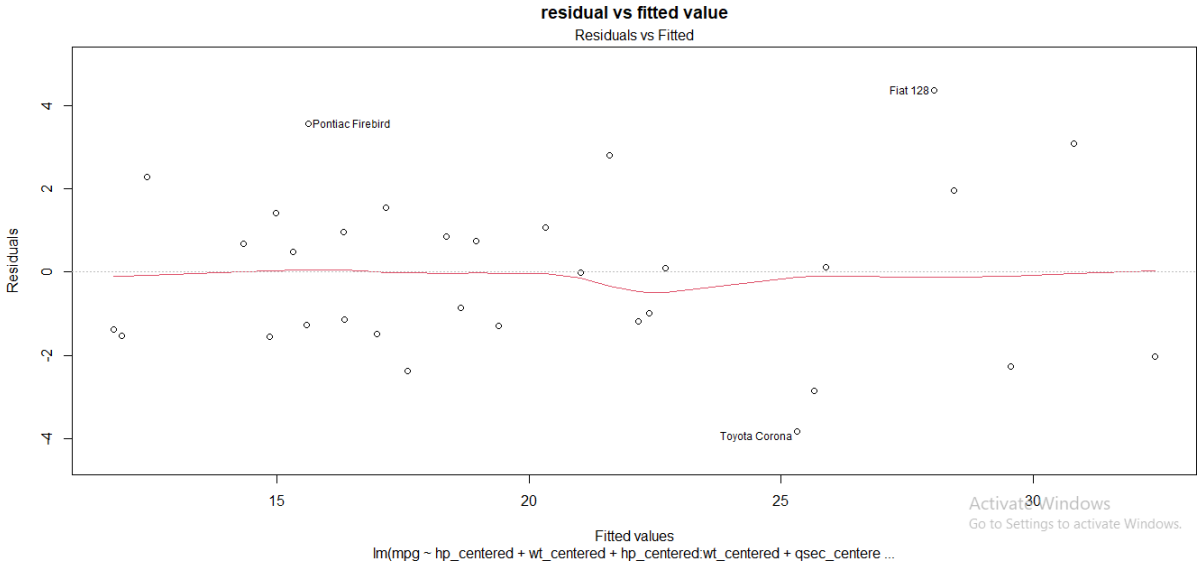
-----Here is our code-----

####Rechecking the assumption linearity###

###by using residual vs fitted value plot###

plot(model_reduced,which=1,main="residual vs fitted value")

output:



Interpretation of Linearity

The **Residual vs. Fitted plot** helps assess whether the assumption of linearity holds in a regression model.

- **New Model (Improved Model) Observations**:
  - The residuals are randomly scattered around **zero**, without a distinct pattern or curve.
  - The red **smoother line** (loess curve) is relatively flat, indicating a good linear fit.

- **Comparison with the Original Model:**
  - In the **original model**, there might have been **curvature**, suggesting **non-linearity** between predictors and mpg.
  - The **new model** addresses this by adding interaction terms and removing multicollinear variables, leading to a more linear relationship.

- **Conclusion:** The assumption of **linearity is better satisfied** in the improved model compared to the original.

Recheck the assumption Homoscedasticity

-----Here is our code-----

```
####Recheck constant variance(Homoscedasticity) assumption###

    ### by using goldfeld-quandt test##

gqtest(model_reduced, order.by = ~ hp, data = mtcars, fraction = 0.2)
```

output:

```
        Goldfeld-Quandt test

data:  model_reduced
GQ = 0.54167, df1 = 8, df2 = 7, p-value = 0.7956
alternative hypothesis: variance increases from segment 1 to 2
```

Interpretation of the Goldfeld-Quandt Test

The **Goldfeld-Quandt test** is used to check for **heteroscedasticity** (i.e., non-constant variance of residuals).

- **Observations from the Output:**
  - **Test Statistic (GQ):** 0.54167
  - **Degrees of Freedom:** df1 = 8, df2 = 7
  - **p-value: 0.7956**

- **Interpretation:**
  - The **null hypothesis (H_0)** assumes **homoscedasticity** (constant variance).
  - The **alternative hypothesis (H_1)** assumes **heteroscedasticity** (variance increases from segment 1 to 2).
  - Since **p-value (0.7956) > 0.05**, we **fail to reject H_0** , meaning there is no significant evidence of heteroscedasticity.
- **Conclusion:** Homoscedasticity assumption is satisfied in the reduced model. The variance of residuals remains constant, making the model reliable.

ReChecking the assumption of Normality

---Here is our code---

######Recheck assumption of normality###

##by using shapiro wilk test###

shapiro.test(residuals(model_reduced))

output:

```
        Shapiro-Wilk normality test

data:  residuals(model_reduced)
W = 0.97535, p-value = 0.6578
```

Interpretation of the Shapiro-Wilk Normality Test

The Shapiro-Wilk test is used to assess whether a dataset follows a normal distribution.

- **Observations from the Output:**
  - **Test Statistic (W):** 0.97535
  - **p-value:** 0.6578
- **Interpretation:**
  - The **null hypothesis (H_0)** assumes that the residuals follow a normal distribution.
  - The **alternative hypothesis (H_1)** assumes that the residuals deviate from normality.
  - Since the **p-value (0.6578) > 0.05**, we **fail to reject H_0** , meaning there is no significant evidence to suggest that the residuals are not normally distributed.
- **Conclusion:**The assumption of normality is satisfied in the reduced model. The residuals appear to follow a normal distribution, which supports the validity of the model's statistical inferences.

Recheck the multicollinarity

---Here is our code---

check the assumption of multicolliniarity###

by using variance inflation factor####

vif(model_reduced)

output:

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
            hp_centered             wt_centered           qsec_centered
               4.924442                2.558084                2.873893
hp_centered:wt_centered
               1.019414
```

Interpretation of Multicollinearity (VIF) in the Reduced Model

Variance Inflation Factor (**VIF**) is used to detect multicollinearity among predictors in a regression model. Lower VIF values indicate reduced collinearity, improving model stability.

- **New Model (Improved Model) Observations:**
  - The VIF values for all predictors are significantly lower compared to the original model, meaning multicollinearity has been reduced.
  - The interaction term **(`hp_centered:wt_centered`)** has a very low VIF **(1.02)**, indicating it does not introduce multicollinearity.
  - The highest VIF in the reduced model is **hp_centered (4.92)**, which is below the critical threshold of concern (typically **VIF > 5 or 10**).

- **Comparison with the Original Model:**
  - In the original model, **`wt`**, **`cyl`**, and **`disp`** had high VIF values (7.17, 7.80, and 10.41, respectively), indicating severe multicollinearity**.**
  - The new model addresses this issue by centering variables, adding interaction terms, and removing collinear predictors (**`cyl`** and **`disp`)**, leading to a more stable and interpretable model.
  - The VIF of **`wt`** significantly dropped from 7.17 to 2.56, and `qsec` also improved (3.63 → 2.87).

- **Conclusion:**

  - The assumption of low multicollinearity is better satisfied in the improved model compared to the original.

  - Removing multicollinear variables and adding interaction terms resulted in a more stable regression model.

  - The reduced model is more reliable and provides more accurate coefficient estimates.

SUMMARY OF REDUCED MODEL

---Here is our code---

###Display the summary of reduced model###

summary(model_reduced)

output:

```
Call:
lm(formula = mpg ~ hp_centered + wt_centered + hp_centered:wt_centered +
    qsec_centered, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8243 -1.3980  0.0303  1.1582  4.3650

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             18.900833   0.487545  38.767  < 2e-16 ***
hp_centered             -0.016769   0.012308  -1.362 0.184320
wt_centered             -4.604876   0.621612  -7.408 5.72e-08 ***
qsec_centered            0.503163   0.360768   1.395 0.174476
hp_centered:wt_centered  0.027791   0.007298   3.808 0.000733 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.117 on 27 degrees of freedom
Multiple R-squared:  0.8925,    Adjusted R-squared:  0.8766
F-statistic: 56.05 on 4 and 27 DF,  p-value: 1.094e-12
```

INTERPRITATION:

This is the output of a new(improved) linear regression model with `mpg` as the dependent variable, using **centered** versions of `hp` (horsepower) and `wt` (weight), along with their interaction term (`hp_centered:wt_centered`), and `qsec` as predictors. Let's break down and compare the results:

1. Residuals:

- The range of residuals is from -3.8243 to 4.3650. This is a slightly narrower range compared to the original model, suggesting that the predictions may be a little closer to the actual values in this model.

2. Coefficients:

- **Intercept (9.92):**
  - This is the expected value of `mpg` when all predictor variables (`hp_centered`, `wt_centered`, and `qsec`) are 0.

- **hp_centered (-0.01677):**
  - The coefficient for `hp_centered` suggests that for each unit increase in centered horsepower, the `mpg` decreases by about 0.01677.
- **wt_centered (-4.60488):**
  - The coefficient for `wt_centered` indicates that for each unit increase in centered weight, `mpg` decreases by approximately 4.60.
- **qsec (0.50316):**
  - The coefficient for `qsec` suggests that for each unit increase in quarter-mile time, `mpg` increases by approximately 0.50.
- **hp_centered:wt_centered (0.02779):**
  - The interaction term between centered horsepower (`hp_centered`) and centered weight (`wt_centered`) has a coefficient of 0.02779. This means that the relationship between horsepower and `mpg` changes depending on the weight of the car. Specifically, as weight increases, the negative effect of horsepower on `mpg` becomes slightly less severe.

3. Model Fit:

- **Residual standard error**: 2.117, which is lower than the 2.547 from the original model. This indicates that the reduced model has slightly better predictive accuracy.
- **Multiple R-squared**: 0.8925, which means the model explains about 89.25% of the variation in `mpg`. This is an improvement over the original model, where the R-squared was 0.8502.
- **Adjusted R-squared**: 0.8766, which is also higher than the original model's adjusted R-squared of 0.8214, suggesting that the reduced model is a better fit even after adjusting for the number of predictors.
- **F-statistic**: 56.05 with a p-value of 1.094e-12, indicating that the model as a whole is statistically significant. This is a much stronger result than the original model's F-statistic of 29.51, suggesting the reduced model is more powerful.

Comparison:

- **Model Fit**: The reduced model (with centered variables and the interaction term) has a higher R-squared and adjusted R-squared, indicating a better fit to the data than the original model.
- **Significant Predictors**: The reduced model introduces a significant interaction term (`hp_centered:wt_centered`), which was absent in the original model. This suggests that the effect of horsepower on `mpg` depends on the weight of the car. Additionally, `wt_centered` is more significant in the reduced model than in the original.
- **Residuals**: The residuals in the reduced model are a bit smaller, suggesting better overall model performance.

In summary: The reduced model offers better predictive accuracy and fit, with a lower residual standard error and higher R-squared values. The significant interaction between horsepower and weight reveals a more nuanced relationship with `mpg`.