Session 4: Regression

Tingfang Wang

June 2024

Contents

Introduction **1**

mtcars Dataset **1**

Scatterplot **3**

Correlation **4**

Correlation Test **5**

Fitting the Regression Line **5**

Plotting the Regression Line **6**

Testing Significance of Coefficients **6**

Conclusion **7**

Introduction

In this analysis, we will explore a simple regression using the `mtcars` dataset, which contains fuel economy data for various car models. The goal is to examine the relationship between the `miles per gallon (mpg)` and the `horsepower (hp)` of the cars.

mtcars Dataset

The `mtcars` dataset in R contains information about various car models. It is a built-in dataset that comes with the base R installation, so you don't need to install any additional packages to access it. Here's an explanation of the columns present in the mtcars dataset:

- `mpg`: Miles per gallon. This column represents the number of miles a car can travel per gallon of fuel. Higher values indicate better fuel efficiency.

- `cyl`: Number of cylinders. This column represents the number of cylinders in the car's engine. It can take values of 4, 6, or 8, indicating the different engine configurations.
- `disp`: Displacement (in cubic inches). This column represents the total volume of all the cylinders in the engine. It is a measure of the engine's capacity or size.
- `hp`: Horsepower. This column represents the power output of the engine. It is a measure of the engine's performance and capability.
- `drat`: Rear axle ratio. This column represents the ratio of the number of revolutions the driveshaft makes compared to the rear axle for one complete rotation of the wheels. It is related to the car's acceleration and top speed.
- `wt`: Weight (in 1000 pounds). This column represents the weight of the car. It is a measure of the car's mass and can influence its performance and fuel efficiency.
- `qsec`: 1/4 mile time. This column represents the time it takes for the car to cover a quarter-mile distance from a standing start. It is related to the car's acceleration capabilities.
- `vs`: Engine type (0 = V-shaped, 1 = straight). This column represents the type of engine, either V-shaped or straight. It is a categorical variable.
- `am`: Transmission type (0 = automatic, 1 = manual). This column represents the type of transmission, either automatic or manual. It is a categorical variable.
- `gear`: Number of forward gears. This column represents the number of gears in the car's transmission. It is a discrete variable that can take values from 3 to 5.
- `carb`: Number of carburetors. This column represents the number of carburetors in the engine. Carburetors are responsible for mixing air and fuel in internal combustion engines.

```r
# Access the mpg dataset
data(mtcars)

# Display the structure of the dataset
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```r
# View the first few rows of the dataset
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
```

Scatterplot

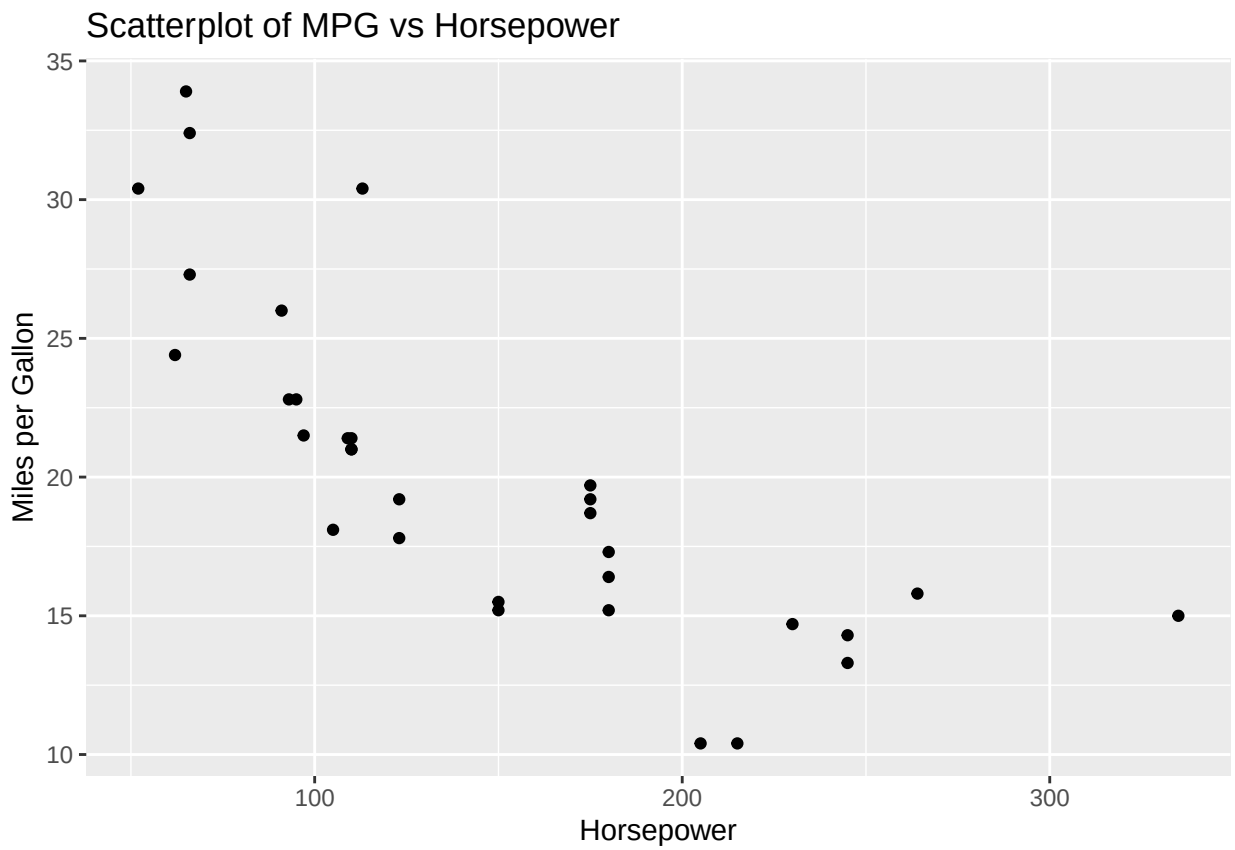
Scatterplots are useful for visualizing the relationship between two numeric variables. The ggplot2 package provides functions to create scatterplots.

To visualize the relationship between `mpg` and `hp`, let's create a scatterplot using ggplot2.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
ggplot(data = mtcars, aes(x = hp, y = mpg)) +
  geom_point() +
  labs(x = "Horsepower", y = "Miles per Gallon") +
  ggtitle("Scatterplot of MPG vs Horsepower")
```

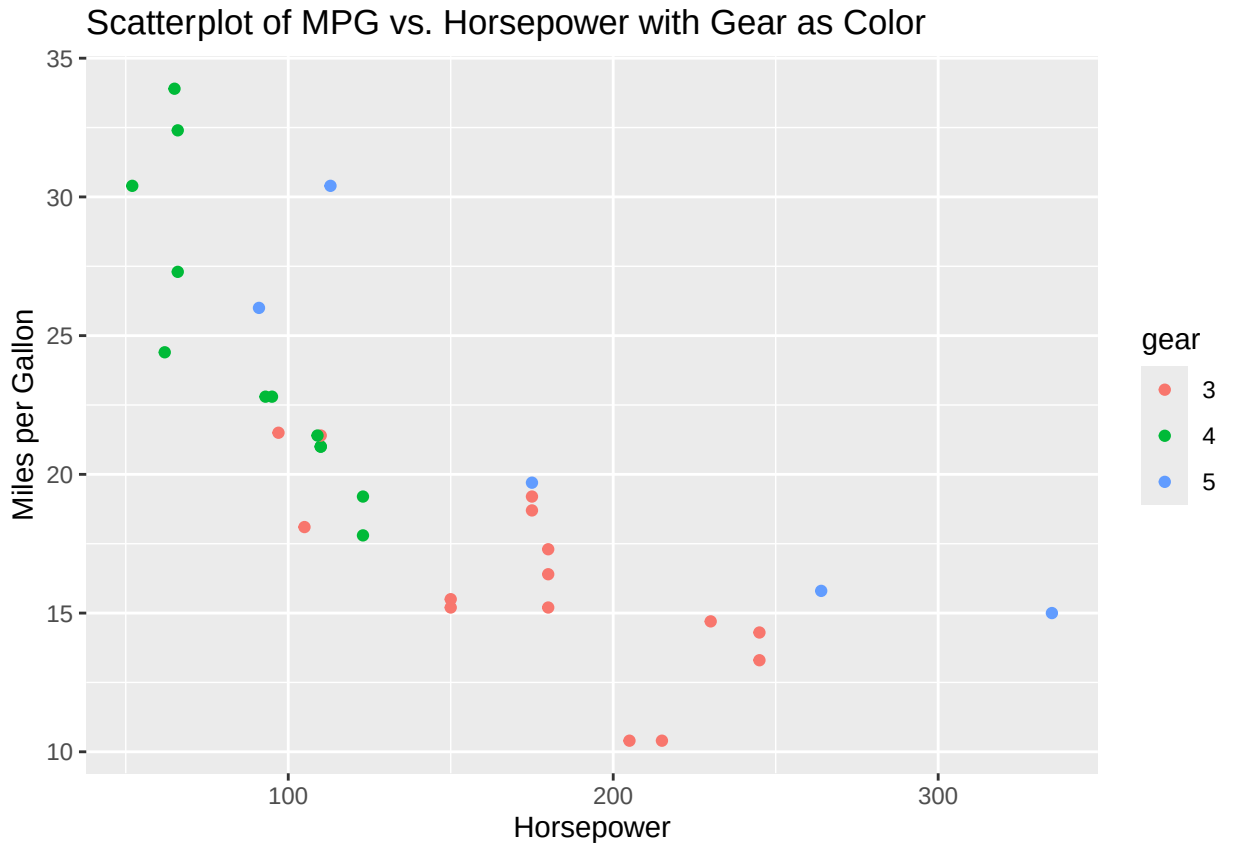


The scatterplot displays the relationship between mpg and hp. We can observe that there appears to be a negative linear relationship between the two variables.

Further, we can map the colors of the points to the 'gear' variable to reveal the number of forward gears of each car.

```r
mtcars$gear <- as.factor(mtcars$gear)
# Create a scatterplot
scatterplot <- ggplot(data = mtcars, aes(x = hp, y = mpg, color = gear)) +
  geom_point() +
  labs(x = "Horsepower", y = "Miles per Gallon") +
  ggtitle("Scatterplot of MPG vs. Horsepower with Gear as Color")

# Print the scatterplot
print(scatterplot)
```



In the code above, we specify mpg as the response variable (y-axis) and hp as the predictor variable (x-axis). The gear column is used as a categorical variable to assign different colors to the points on the scatterplot. By converting gear to a factor using factor(gear), we ensure that the different gear values are treated as distinct categories and displayed with different colors on the scatterplot.

Correlation

To assess the strength and direction of the relationship between `mpg` and `hp`, let's calculate the correlation coefficient.

```r
correlation <- cor(mtcars$mpg, mtcars$hp)
correlation
```

```
## [1] -0.7761684
```

The correlation coefficient between mpg and hp is -0.7761684. This indicates a moderate negative correlation between the two variables.

Correlation Test

To determine if the observed correlation is statistically significant, we can perform a correlation test.

```
cor_test <- cor.test(mtcars$mpg, mtcars$hp)
cor_test
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$mpg and mtcars$hp
## t = -6.7424, df = 30, p-value = 1.788e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8852686 -0.5860994
## sample estimates:
##        cor
## -0.7761684
```

The correlation test reveals that the correlation between mpg and hp is statistically significant (p-value = $1.7878353 \times 10^{-7} < 0.05$).

Fitting the Regression Line

To quantify the relationship between `mpg` and `hp`, let's fit a simple linear regression model to the data.

```
model <- lm(mpg ~ hp, data = mtcars)
summary(model)
```

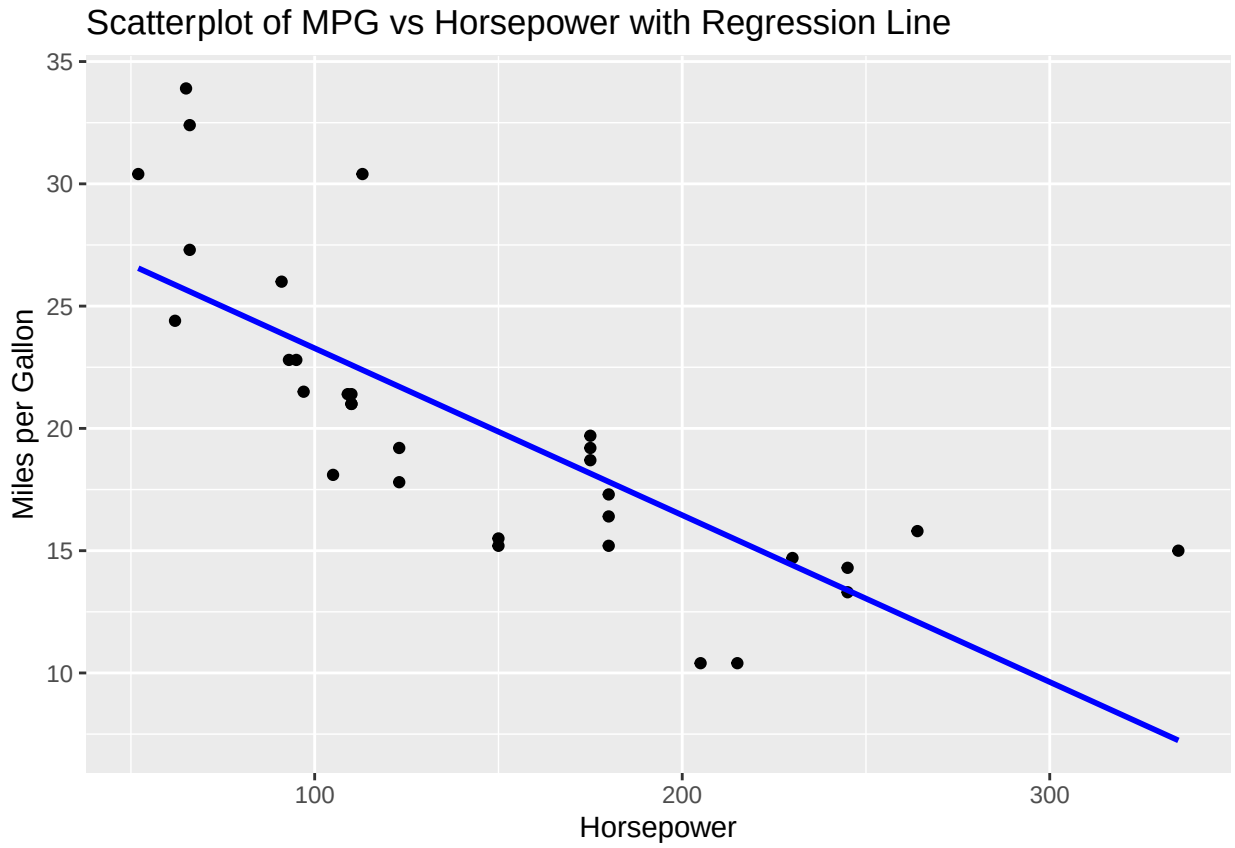
```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

The regression model indicates that the equation for the fitted line is $\text{mpg} = 30.10 - 0.07 * \text{hp}$, where 30.10 is the intercept and - 0.07 is the coefficient for hp. The coefficient for hp is statistically significant (p-value $= 1.7878353 \times 10^{-7} < 0.05$), suggesting a significant relationship between mpg and hp.

Plotting the Regression Line

To visualize the regression line on the scatterplot, let's add it to the existing plot.

```
ggplot(data = mtcars, aes(x = hp, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "blue") +
  labs(x = "Horsepower", y = "Miles per Gallon") +
  ggtitle("Scatterplot of MPG vs Horsepower with Regression Line")
```



The scatterplot now includes the regression line, which represents the estimated relationship between mpg and hp.

Testing Significance of Coefficients

To test the significance of the coefficients in the regression model, we can use hypothesis tests.

```r
coefTests <- coef(summary(model))
coefTests
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 30.09886054  1.6339210 18.421246 6.642736e-18
## hp          -0.06822828  0.0101193 -6.742389 1.787835e-07
```

The table above displays the coefficients, standard errors, t-values, and p-values for each predictor variable. The p-value for the hp coefficient is less than 0.05, indicating that it is statistically significant.

Conclusion

In this analysis, we explored the relationship between mpg and hp using simple linear regression and ggplot2 for visualization. The scatterplot and regression analysis revealed a negative relationship between the two variables. The correlation test and coefficient tests confirmed the statistical significance of this relationship.

Please note that this analysis assumes a linear relationship between mpg and hp and does not account for other potential confounding variables.