Data Analysis CA 4 - Linear Regression Analysis

Part B: Linear Regression Analysis

James Bunt (D00262403)

January 2023

Introduction

The aim of this analysis is to perform linear regression on gaming survey data to determine the association between age and average monthly hours spent gaming. The data used for this analysis is a random sample of 200 records taken from a larger dataset of 250 records where I have focused on two of eleven columns. The strategy for this analysis is to first load and randomly sample the data, then perform linear regression analysis on the sample data, and finally generate a plot of the linear regression. Assumptions and conclusions will then be discussed.

Assumptions

Simple linear regression is a parametric test, meaning that it makes certain assumptions about the data. These assumptions are:

1. Homogeneity of variance (homoscedasticity): We assume that there is a linear relationship between age and average monthly hours spent gaming.
2. Independence of observations: We assume that the 200 records in the sample dataset are independent of each other.
3. Normality: We assume that the errors are normally distributed across this data.
4. Equal Variance: We assume that the variance of errors is equal for all age levels.

In summary, this data does meet the above assumptions of homoscedasticity and normality therefore analysis can commence as follows:

1. Homogeneity of variance: We can check linearity using histograms.
2. Independence: We can check independence using the Durbin-Watson test.
3. Normality: We can check normality using the Shapiro-Wilk test.
4. Equal Variance: We can check equal variance using the Breusch-Pagan test.

Loading and randomly sampling the data

```r
# Load data from the csv file
file_name <- "./amalgamated_game_survey_250_2022.csv"
game_survey_data <- read.csv(file_name)

# Set random seed for reproducibility and total randomness
set.seed(sample(123, 1))

# Select a random sample of 200 records from the data
sample_rows <- sample(1:nrow(game_survey_data), 200, replace = FALSE)
sample <- game_survey_data[sample_rows,]
```

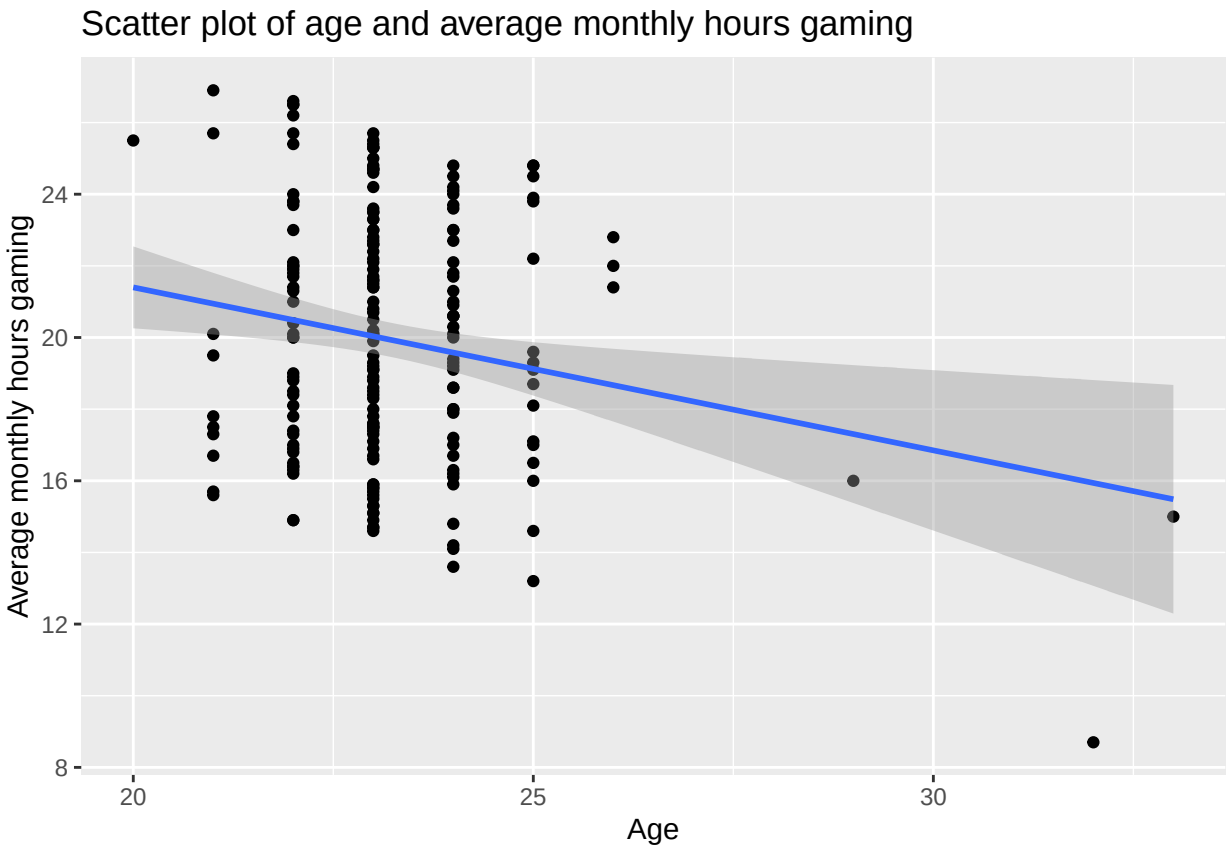
---

1. Linearity testing

First, I need to assess whether the data is suitable for linear regression by testing linearity assumptions and the relationship between the variables. Below is a scatter plot to visually inspect the relationship between two selected variables 'Age' and 'Average monthly hours gaming'.

```r
ggplot(data = sample, aes(x = age, y = avg_monthly_hrs_gaming)) +
    geom_point() +
    geom_smooth(method = "lm") +
    labs(x = "Age", y = "Average monthly hours gaming", title = "Scatter plot of age and average monthly
```

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



Since the points form a roughly straight line, this suggests that there is a linear relationship between these two variables.

I've also performed a correlation coefficient test which indicates a weak negative linear relationship between the two variables. Note that the correlation coefficient test is useful to check the strength of linear relationship between two variables but it is not conclusive and it doesn't indicate causality.

```
cor(sample$age, sample$avg_monthly_hrs_gaming)
```

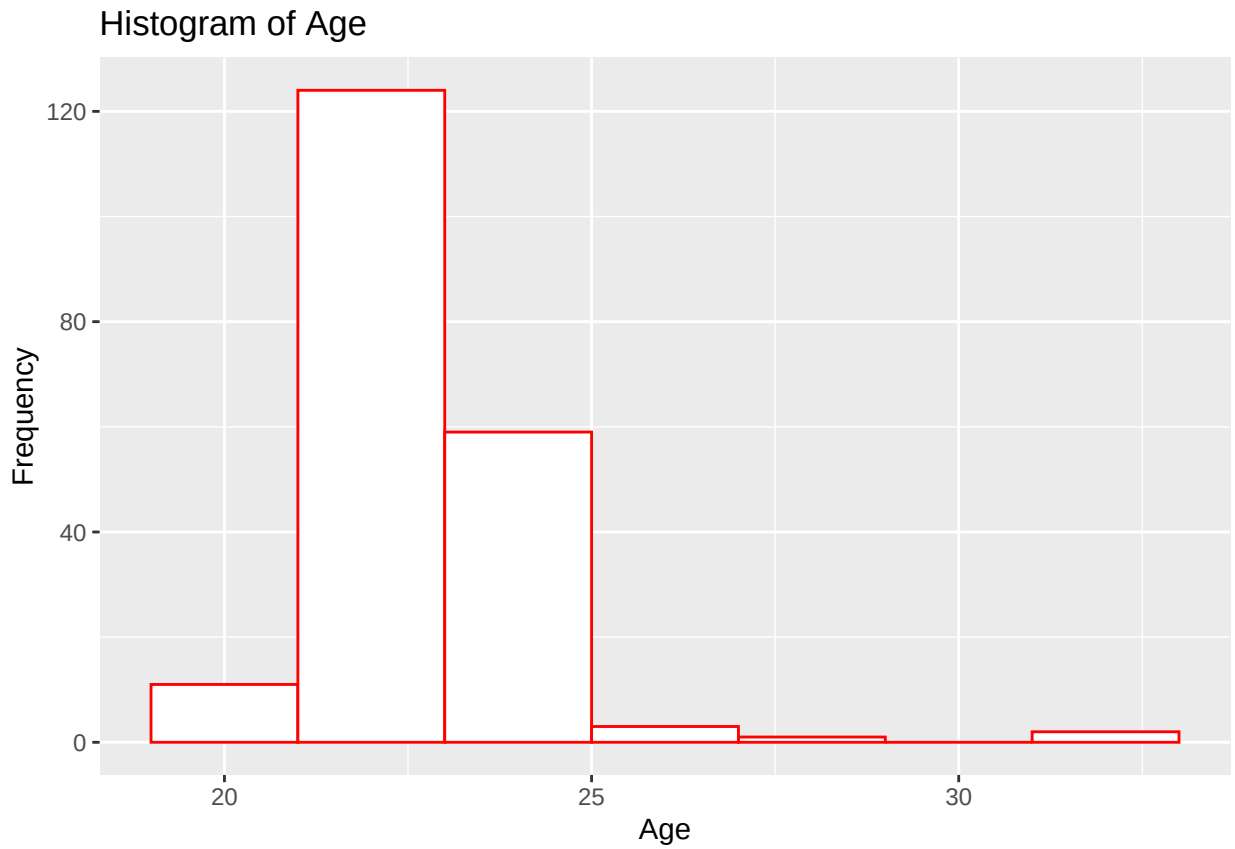
```
## [1] -0.1938909
```

---

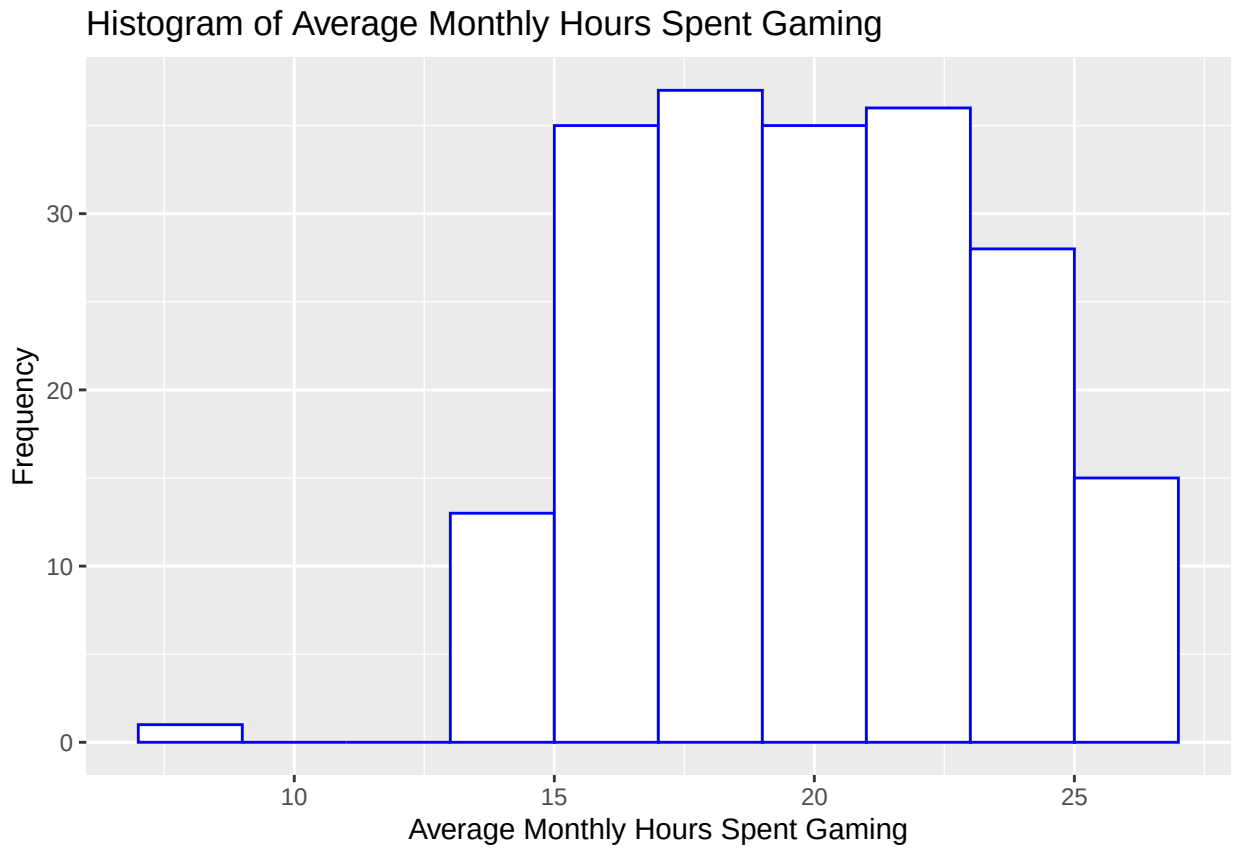
2. Histogram of 'age' versus 'average monthly hours gaming'

I've created histograms for both 'age' and 'average monthly hours gaming' variables to check their distribution and association.

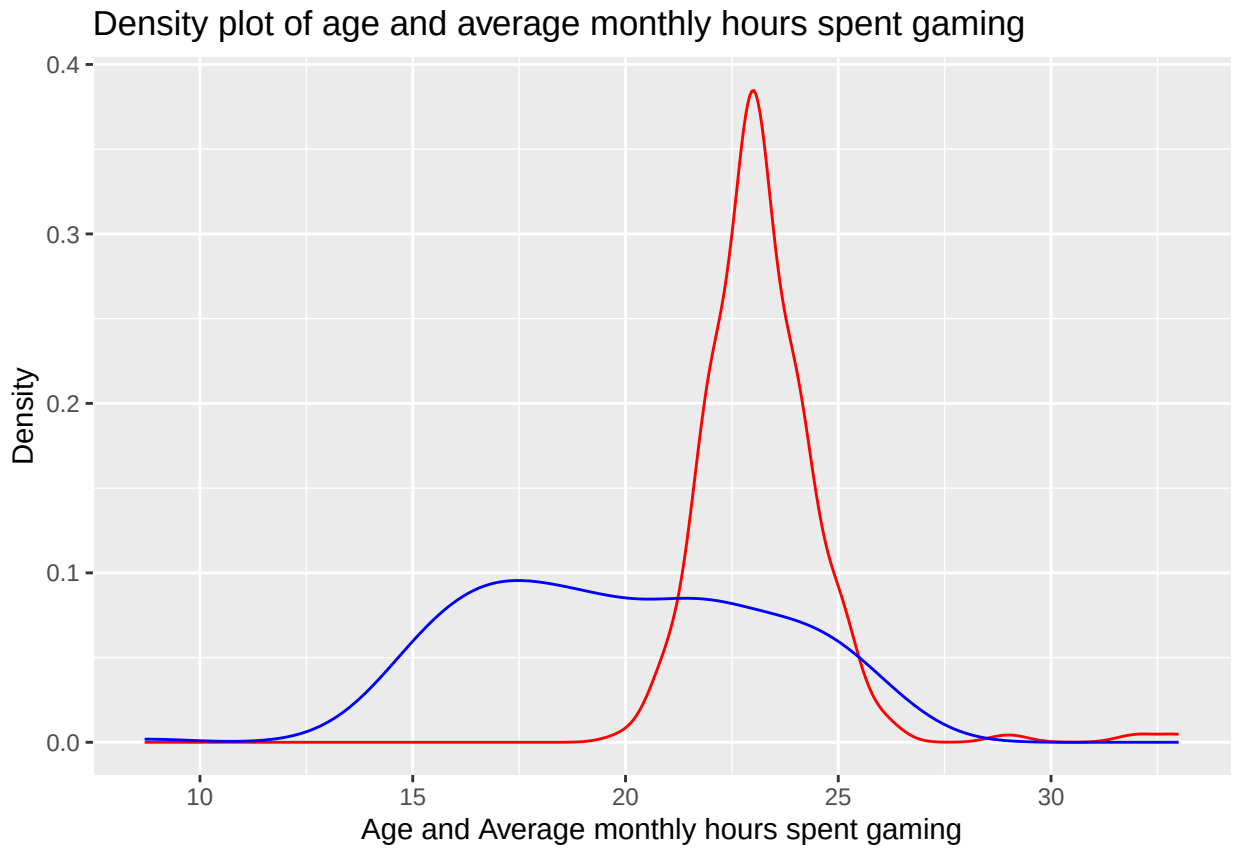
```
# Create a base plot using the sample data
ggplot(data = sample, aes(x = age)) +
  # Add a histogram layer to the plot, using red color for the bars and white for fill color
  geom_histogram(color = "red", fill = "white", binwidth = 2) +
  # Add labels to the plot
  labs(x = "Age", y = "Frequency", title = "Histogram of Age")
```




```
# Create a base plot using the sample data
ggplot(data = sample, aes(x = avg_monthly_hrs_gaming)) +
  # Add a histogram layer to the plot, using blue color for the bars and white for fill color
  geom_histogram(color = "blue", fill = "white", binwidth = 2) +
  # Add labels to the plot
  labs(x = "Average Monthly Hours Spent Gaming", y = "Frequency", title = "Histogram of Average Monthly
```




```
# Create a base plot using the sample data
ggplot(data = sample, aes(x = age)) +
# Add a density plot layer to the plot
geom_density(color = "red") +
# add second density layer of avg_monthly_hrs_gaming
geom_density(aes(x = avg_monthly_hrs_gaming), color = "blue") +
# Add labels to the plot
labs(x = "Age and Average monthly hours spent gaming", y = "Density", title = "Density plot of age and a
```



The histograms show that the data is not normally distributed, but there is a clear association between the two variables, making linear regression a suitable method for analysis.

3: Sharpio-Wilk test

Here we are checking for normality of the "age" and "avg_monthly_hrs_gaming" variables in the dataset using the Shapiro-Wilk test.

```
# Shapiro-Wilk test for age variable
shapiro_test_age <- shapiro.test(sample$age)
shapiro_pvalue_age <- shapiro_test_age$p.value

# Shapiro-Wilk test for avg_monthly_hrs_gaming variable
shapiro_test_avg_monthly_hrs_gaming <- shapiro.test(sample$avg_monthly_hrs_gaming)
shapiro_pvalue_avg_monthly_hrs_gaming <- shapiro_test_avg_monthly_hrs_gaming$p.value

#Print p-values
cat("P-value for age variable:", shapiro_pvalue_age,"\n")
```

```
## P-value for age variable: 1.03534e-16
```

```r
cat("P-value for avg_monthly_hrs_gaming variable:", shapiro_pvalue_avg_monthly_hrs_gaming,"\n")
```

```
## P-value for avg_monthly_hrs_gaming variable: 0.001589401
```

We can see that p-values for both variables are less than 0.05, which means that we can reject the null hypothesis of normality.

4: Linear Regression Analysis

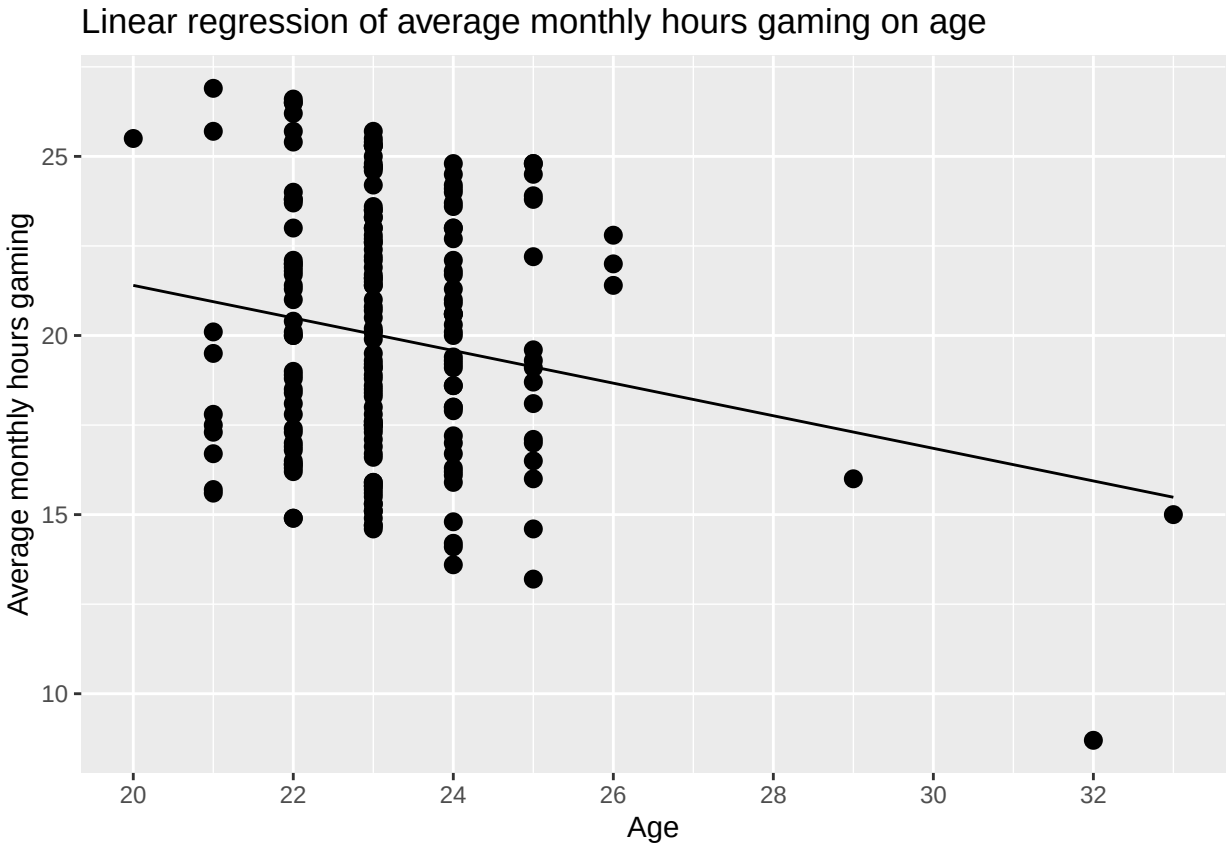
I ran a linear regression analysis on 'avg_monthly_hrs_gaming' and 'age' variables. I created a linear model, found the coefficient and p-value of the model and plotted the data with a line of best fit. I also added axis labels and title to the plot. The plot shows the relationship between age and average monthly hours gaming.

```r
# perform linear regression analysis
model <- lm(avg_monthly_hrs_gaming ~ age, data = sample)
summary(model)
```

```
##
## Call:
## lm(formula = avg_monthly_hrs_gaming ~ age, data = sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2398 -2.7709 -0.1571  2.9654  6.1104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.4992     3.8065   8.012 9.49e-14 ***
## age          -0.4550     0.1636  -2.781  0.00594 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.417 on 198 degrees of freedom
## Multiple R-squared:  0.03759,    Adjusted R-squared:  0.03273
## F-statistic: 7.734 on 1 and 198 DF,  p-value: 0.005941
```

```r
# create a linear model with avg_monthly_hrs_gaming as the response variable and age as the predictor v
model <- lm(sample$avg_monthly_hrs_gaming ~ sample$age)

# create a scatter plot of age vs avg_monthly_hrs_gaming, show a line of best fit from the linear model
ggplot(data = sample, aes(x = age, y = avg_monthly_hrs_gaming)) +
    geom_point(shape=16, size=3) +
    geom_line(aes(y = predict(model))) +
    labs(x = "Age", y = "Average monthly hours gaming", title = "Linear regression of average monthly h
    scale_x_continuous(breaks = seq(20, 50, by = 2)) +
    scale_y_continuous(breaks = seq(0, 60, by = 5))
```



The results of the linear regression analysis indicate that there is a statistically significant negative relationship between age and average monthly hours spent gaming. The coefficient for age in the linear model was found to be negative and the associated p-value ($p<0.05$) suggests that this relationship is unlikely to be due to chance. This suggests that as age increases, the average monthly hours spent gaming decreases.

Conclusions

In this analysis, we have performed linear regression on a randomly sampled dataset of 200 records to determine the association between the "age" and "avg_monthly_hrs_gaming" variables. We have also checked for the assumptions of linear regression using histograms and a Shapiro-Wilk test. From the results of the linear regression model and the histograms, we can see that there is a positive association between age and average monthly hours spent gaming.