CA 4: Part B - Linear Regression Analysis

Analysis of linear regression performed on randomly sampled data from a pair of columns from a survey of gamer preferences. This includes statement and testing of assumptions followed by intrepretation and dicussion of obtained results.

Gareth Burger

January, 2023

Table of Contents

Aim

To perform linear regression on data from a survey of gamer preferences. The variable pair being investigated is ‘age’ and ‘average monthly expenditure DLC’. The aim is to determine whether the two variables are correlated i.e. does a gamer’s age affect the amount of money they spend on DLC per month. If so, the independent age variable will have a linear relationship with the dependent variable of money spent.

Statement of Assumptions

Independent variable: Age
Dependent variable: Average Monthly Expenditure DLC

Linear regression makes certain assumptions concerning the data it is testing. These four assumptions are described below:

1. **Normality** - the data of the dependent variable follows a normal distribution
2. **Linearity** - the relationship between the independent and dependent variable is linear (the line of best fit is a straight line)
3. **Homoscedasticity** - the variance of the error doesn’t change significantly for each value of the independent variable (the data points don’t diverge significantly from the linear line of best fit)
4. **Independence** - observations are independent of each other

Furthermore, it can be assumed that the data is randomly sampled and that no hidden relationships exist among observations in the dataset.

Testing of Assumptions

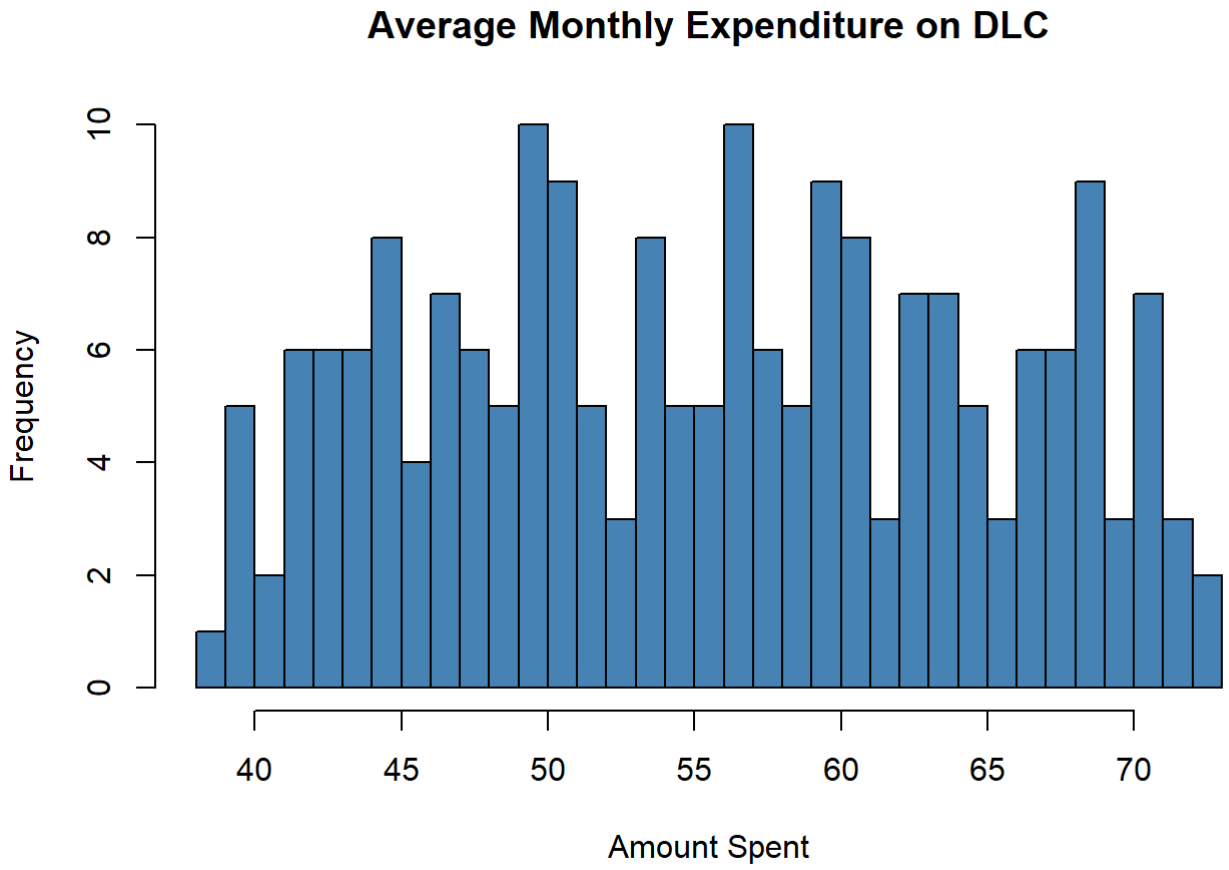
Loading and Randomly Sampling the Data

```
# load data from the csv
file_name = "./amalgamated_game_survey_250_2022.csv"
game_survey_data <- read.csv(file_name)

# sample 200 random elements from the data

# set.seed(some random number) to ensure that we get the same sample each time
set.seed(500)

# sample a random vector of 200 elements from all 250 rows in the game_survey_data dataset
sample_rows <- sample(1:nrow(game_survey_data), 200, replace = FALSE)

# select the 200 rows that match the sampled row numbers
sample <- game_survey_data[sample_rows, ]
```

Testing for Normality

Method 1: Using Histogram / Density Plot

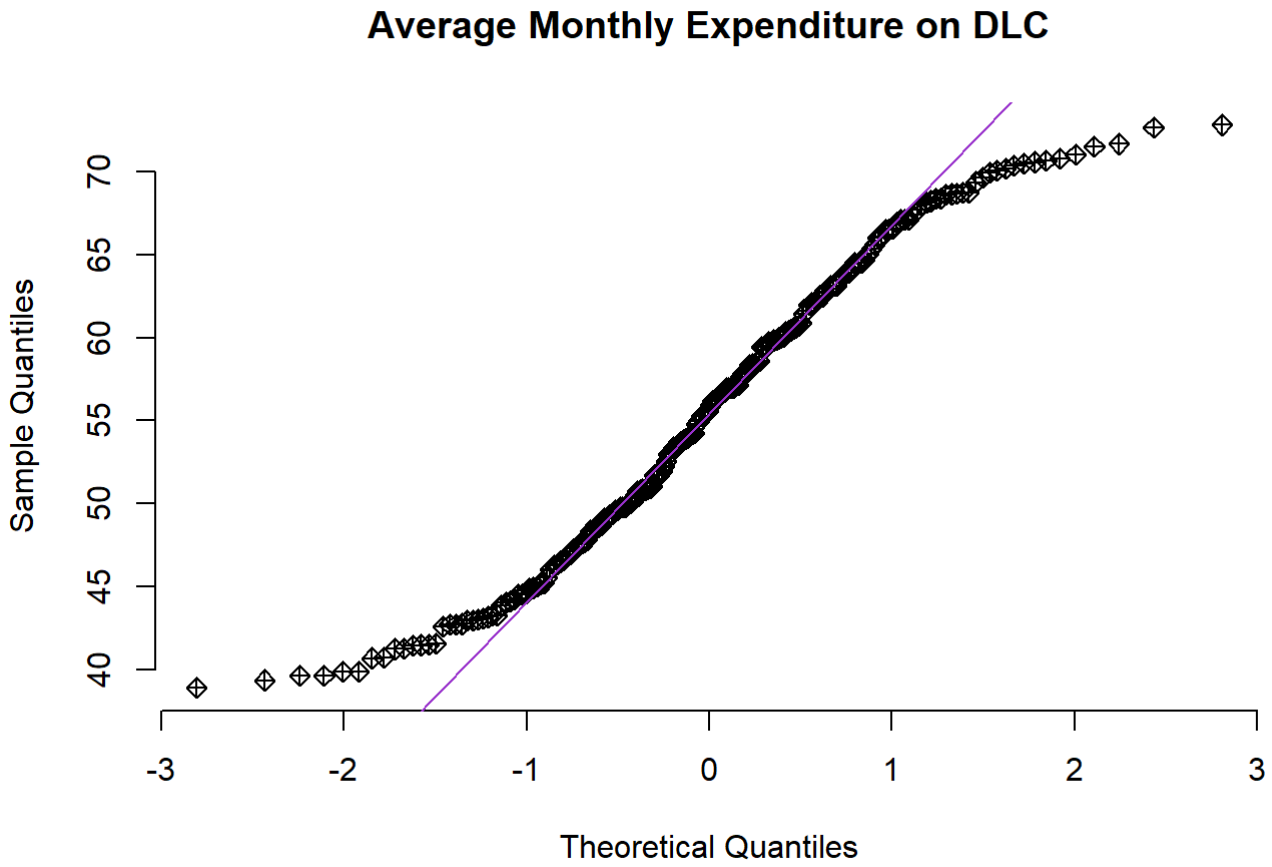
```
# draw histogram for avg_monthly_expenditure_dlc
hist(sample$avg_monthly_expenditure_dlc,
     main = "Average Monthly Expenditure on DLC",
     xlab = "Amount Spent",
     ylab = "Frequency",
     breaks = 30,
     col = 'steelblue')
```



The histogram of the dependent variable does not exhibit data that is normally distributed as it is not bell-shaped in nature, but rather is more of a multimodal distribution.

Method 2: Using Q-Q Plot

```
# perform qqnorm test on avg_monthly_expenditure_dc
qqnorm(sample$avg_monthly_expenditure_dlc,
       pch = 9,
       frame = FALSE,
       main = "Average Monthly Expenditure on DLC")
qqline(sample$avg_monthly_expenditure_dlc,
       col = "darkorchid",
       lwd = 1)
```



In the Q-Q plot above, a large amount of points fall roughly along the diagonal Q-Q line giving us the impression that the data could be normally distributed. However, there are also a significant number of points that do not adhere to the diagonal line, thus further testing for normality is required.

Method 3: Shapiro-Wilk Test

The threshold for normality is 0.05 (5%). If the p-value is above the threshold then we can accept the data to be normally distributed.

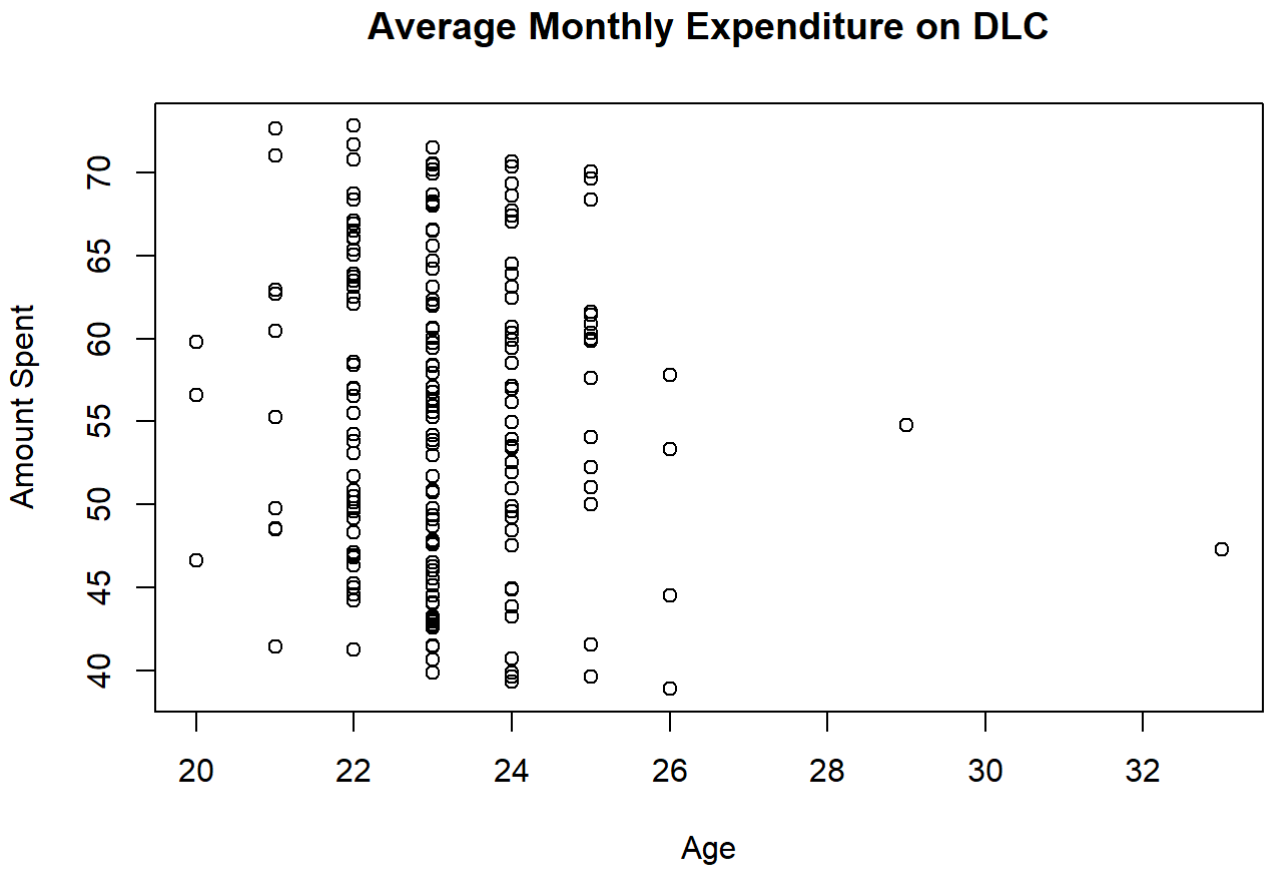
```
# perform Shapiro-Wilk test for normality
st <- shapiro.test(sample$avg_monthly_expenditure_dlc)
st
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample$avg_monthly_expenditure_dlc
## W = 0.96292, p-value = 4.146e-05
```

A Shapiro-Wilk normality test was conducted on the average monthly expenditure DLC sample data.
From the output obtained we cannot assume normality as the p-value (p = 4.146^{-5}) is less than 0.05.

Testing for Linearity

```
# draw a scatter plot for avg_monthly_expenditure_dlc
plot(sample$age, sample$avg_monthly_expenditure_dlc,
     main = "Average Monthly Expenditure on DLC",
     xlab = "Age",
     ylab = "Amount Spent")
```



The distribution of the data points in the scatter plot above cannot be described with a straight line, which indicates that the relationship between the independent and dependent variable is not linear.

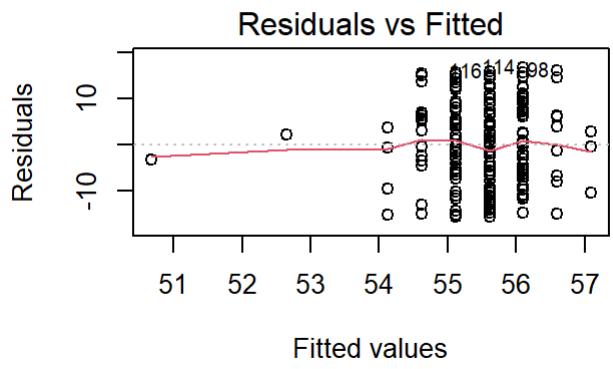
Linear Regression Analysis

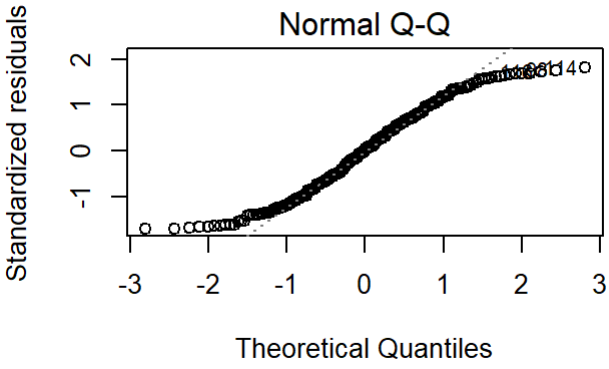
Linear regression analysis is performed to evaluate the relationship between the independent and dependent variables. It also allows us to further test for homoscedasticity (below).

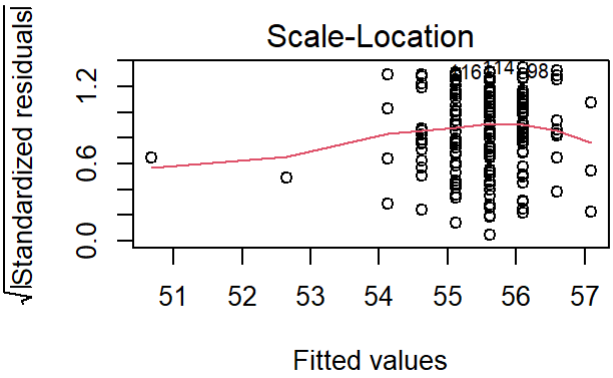
```
# perform simple linear regression analysis on the variable pair and create the model
age_spend_lm <- lm(avg_monthly_expenditure_dlc ~ age, data = sample)
```

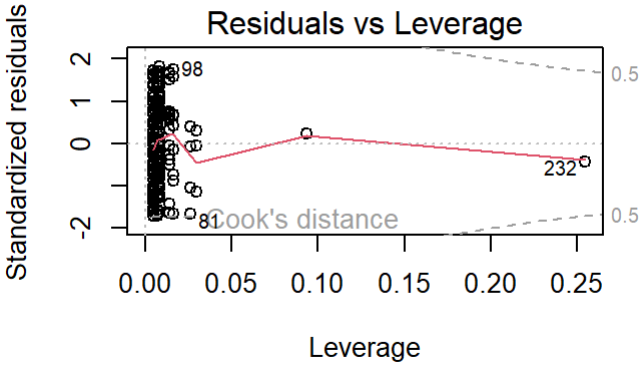
Testing for Homoscedasticity

```
# plot to check for homoscedasticity
par(mfrow=c(2,2))
plot(age_spend_lm)
```










```
par(mfrow=c(1,1))
```

As can be seen in the residual plots above, the mean of the residuals (represented by the red lines) are not horizontal or centered around zero. This means that there are biases in the sample data that make linear regression analysis invalid. The real residuals do not fit with or closely to the theoretical residuals of a perfect model, thus the sample data does not meet the assumption of homoscedasticity.

Results

The results of the assumption testing indicate that the data is not suitable for linear regression, namely:

1. **Normality**
    - The histogram of the dependent variable does not exhibit normally distributed data
    - Sample points of the data do not aptly fit the diagonal Q-Q line in the Q-Q plot
    - The Shapiro-Wilk test results in a p-value that is below the threshold for normally distributed data
2. **Linearity**
    - As seen in the scatter plot, the independent and dependent variable do not have a linear relationship and one is not able to draw a straight line of best fit
3. **Homoscedasticity**
    - There is a significant change in the variance of error for each value of the independent variable and sample points do not fit within straight line bounds, meaning the data is not homogeneous in variance

For the purpose of this exercise, let's assume that the data was suitable for simple linear regression.

```
# display a summary of the linear regression model
summary(age_spend_lm)
```

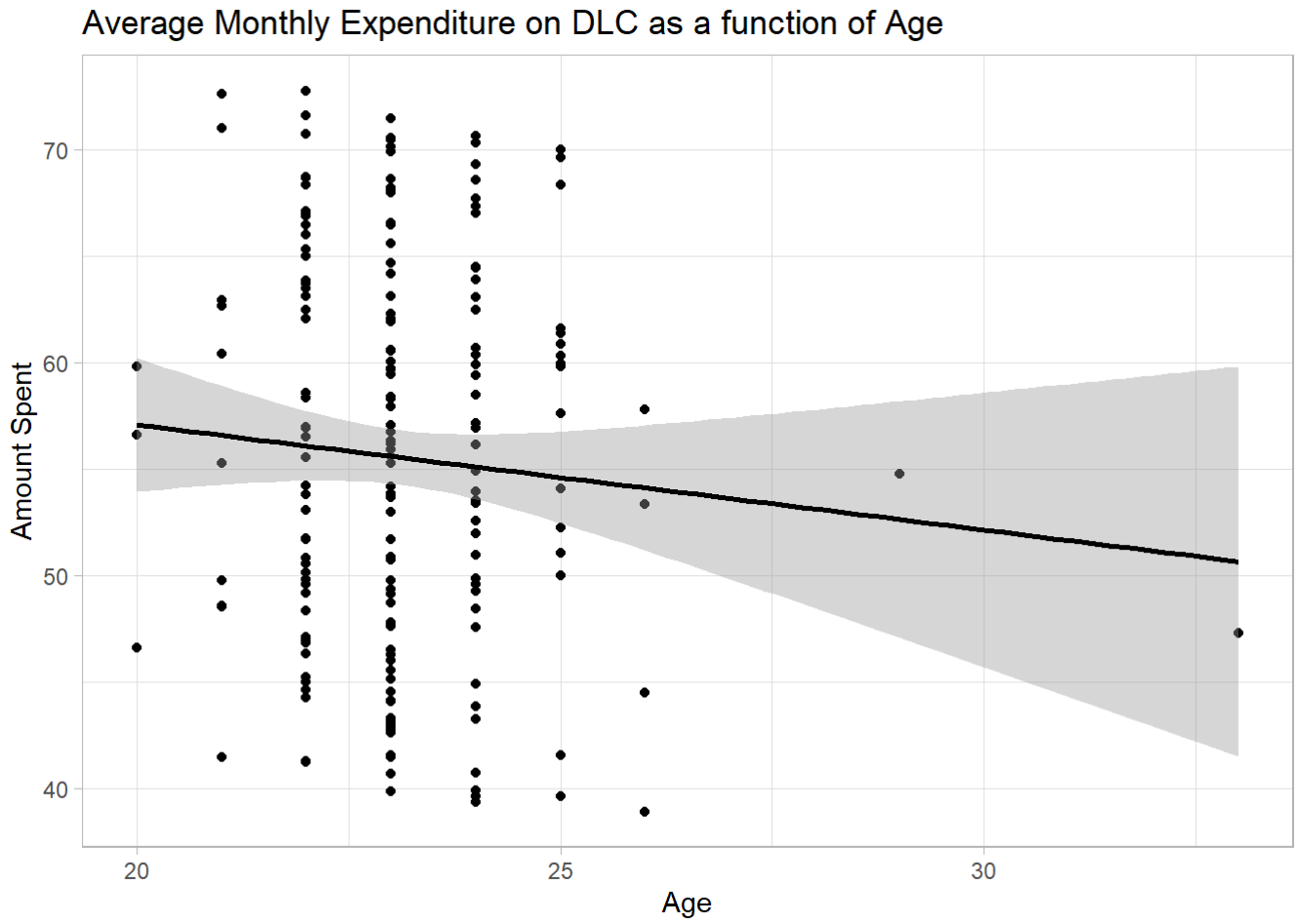
```
##
## Call:
## lm(formula = avg_monthly_expenditure_dlc ~ age, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7278  -7.8117   0.1483   7.3578  16.6743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.9729    10.7904   6.207 3.11e-09 ***
## age          -0.4940     0.4659  -1.060     0.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.213 on 198 degrees of freedom
## Multiple R-squared:  0.005646,   Adjusted R-squared:  0.0006242
## F-statistic: 1.124 on 1 and 198 DF,  p-value: 0.2903
```

In this case, the p-value of the model (p = 0.29) is greater than 0.001 which means we must accept the null hypothesis that there is no statistically significant relationship between age and average monthly expenditure on DLC.

If we were to visualise the results of a valid simple linear regression, the code below plots the data points and linear regression line on a graph:

```
# R code using GGPLOT2 to generate linear regression plot
age_spend_graph <- ggplot(sample, aes(x = age, y = avg_monthly_expenditure_dlc)) +
  labs(x = "Age", y = "Amount Spent", title = "Average Monthly Expenditure on DLC as a functi
on of Age") +
  geom_point() +
  geom_smooth(method="lm", col="black") +
  theme_light()
age_spend_graph
```

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


Average Monthly Expenditure on DLC as a function of Age
Amount Spent
70
60
50
40
20
25
30
Age