Homework#1

Sunny Lee

2021-02-08

This homework assignment is expected to provide a basic understanding of the two-sample t-test, Regression, and Anova. Use the **C2 Games2.csv** data to answer the following questions: In this homework, we would like to understand whether a student could play the game quickly with their right or left hand. We want to compare the completion time with hand used to play. **Make sure to explain every plots and outputs and please do not incude irrelevant information to the question. This homework is due on Monday January 15,2021.**

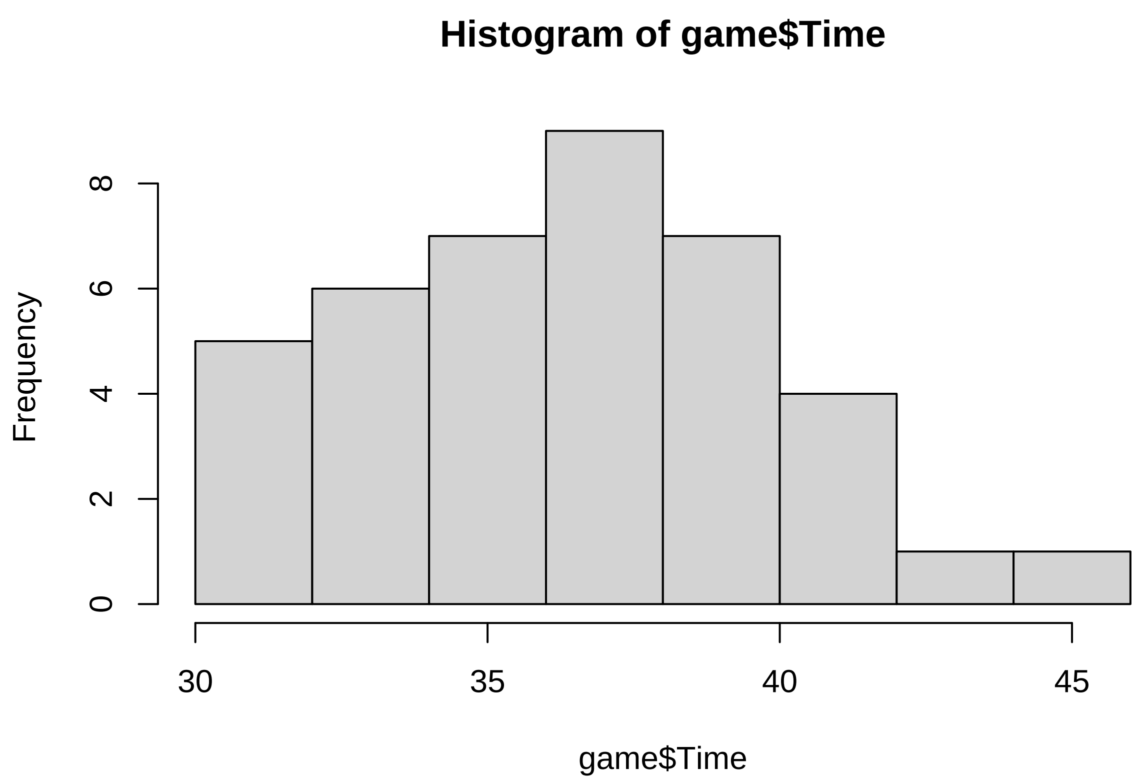
1) Create a histogram and box plot of the completion time for both right and left hand.

```r
game <- read.csv("C2 Games2.csv")
game
```

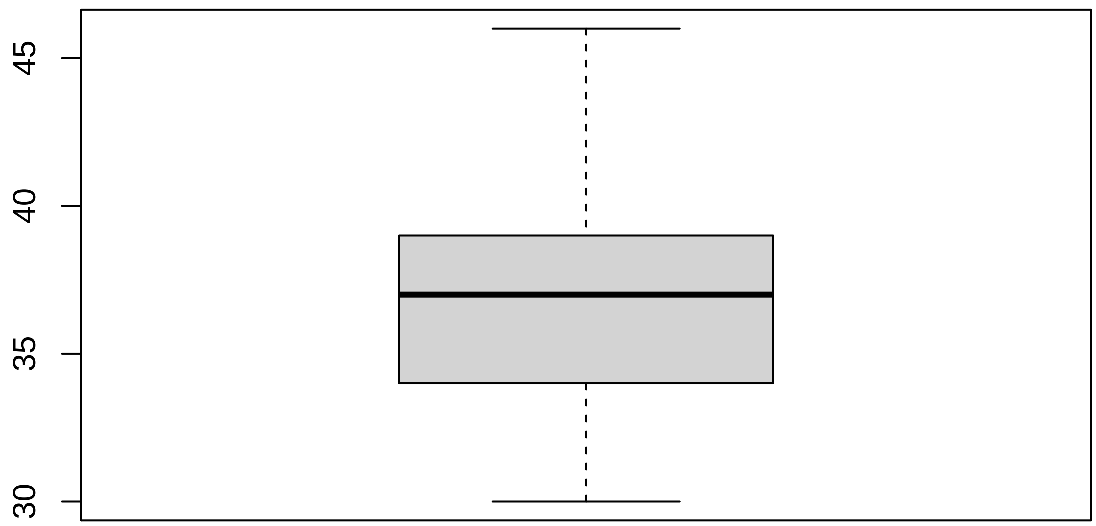
```
##    studentID   numPegs     Type Time  Hand         Type2
## 1          1 twentyone Standard   38 Right StandardRight
## 2          2 twentyone    Color   36 Right    ColorRight
## 3          3 twentyone    Color   42  Left     ColorLeft
## 4          4 twentyone Standard   35 Right StandardRight
## 5          5 twentyone Standard   32  Left  StandardLeft
## 6          6 twentyone    Color   37 Right    ColorRight
## 7          7 twentyone    Color   38  Left     ColorLeft
## 8          8 twentyone Standard   38  Left  StandardLeft
## 9          9 twentyone Standard   37 Right StandardRight
## 10        10 twentyone Standard   36  Left  StandardLeft
## 11        11 twentyone    Color   35  Left     ColorLeft
## 12        12 twentyone    Color   40 Right    ColorRight
## 13        13 twentyone Standard   41  Left  StandardLeft
## 14        14 twentyone Standard   39  Left  StandardLeft
## 15        15 twentyone    Color   33 Right    ColorRight
## 16        16 twentyone    Color   40  Left     ColorLeft
## 17        17 twentyone    Color   46  Left     ColorLeft
## 18        18 twentyone Standard   33 Right StandardRight
## 19        19 twentyone    Color   38 Right    ColorRight
## 20        20 twentyone Standard   31 Right StandardRight
## 21        21 twentyone Standard   36 Right StandardRight
## 22        22 twentyone Standard   37  Left  StandardLeft
## 23        23 twentyone    Color   35 Right    ColorRight
## 24        24 twentyone Standard   33 Right StandardRight
## 25        25 twentyone    Color   37  Left     ColorLeft
## 26        26 twentyone    Color   40  Left     ColorLeft
## 27        27 twentyone    Color   37 Right    ColorRight
## 28        28 twentyone Standard   31  Left  StandardLeft
## 29        29 twentyone Standard   36  Left  StandardLeft
## 30        30 twentyone Standard   34 Right StandardRight
## 31        31 twentyone    Color   34 Right    ColorRight
```

```
## 32         32 twentyone Standard   33 Right StandardRight
## 33         33 twentyone    Color   31 Right    ColorRight
## 34         34 twentyone    Color   44  Left     ColorLeft
## 35         35 twentyone    Color   39  Left     ColorLeft
## 36         36 twentyone Standard   39  Left  StandardLeft
## 37         37 twentyone Standard   42  Left  StandardLeft
## 38         38 twentyone    Color   41  Left     ColorLeft
## 39         39 twentyone    Color   39 Right    ColorRight
## 40         40 twentyone Standard   30 Right StandardRight
```

```
hist(game$Time)
```




```
boxplot(game$Time)
```



From the histogram above, we see the data definitely seems skewed to the left side. The boxplot shows there does not seem to be any outliers.

2) Calculate the mean, sd, and variance completion time for the game played with right hand and left hand.

```
left <- subset(game, game$Hand == "Left")
right <- subset(game, game$Hand == "Right")

mean(right$Time)
```

```
## [1] 35
```

```
sd(right$Time)
```

```
## [1] 2.790963
```

```
var(right$Time)
```

```
## [1] 7.789474
```

From the above, we find that the mean time for the right hand is 35, the sd is 2.790963 and the variance is 7.789474.

```
mean(left$Time)
```

```
## [1] 38.65
```

```
sd(left$Time)
```

```
## [1] 3.674593
```

```r
var(left$Time)
```

```
## [1] 13.50263
```

From the above, we find that the mean time for the right hand is 38.65, the sd is 3.674593 and the variance is 13.50263.

3) Use informal test to determine if the equal variance assumption is appropriate for this study.

```r
max(var(right$Time), var(left$Time)) / min(var(right$Time), var(left$Time))
```

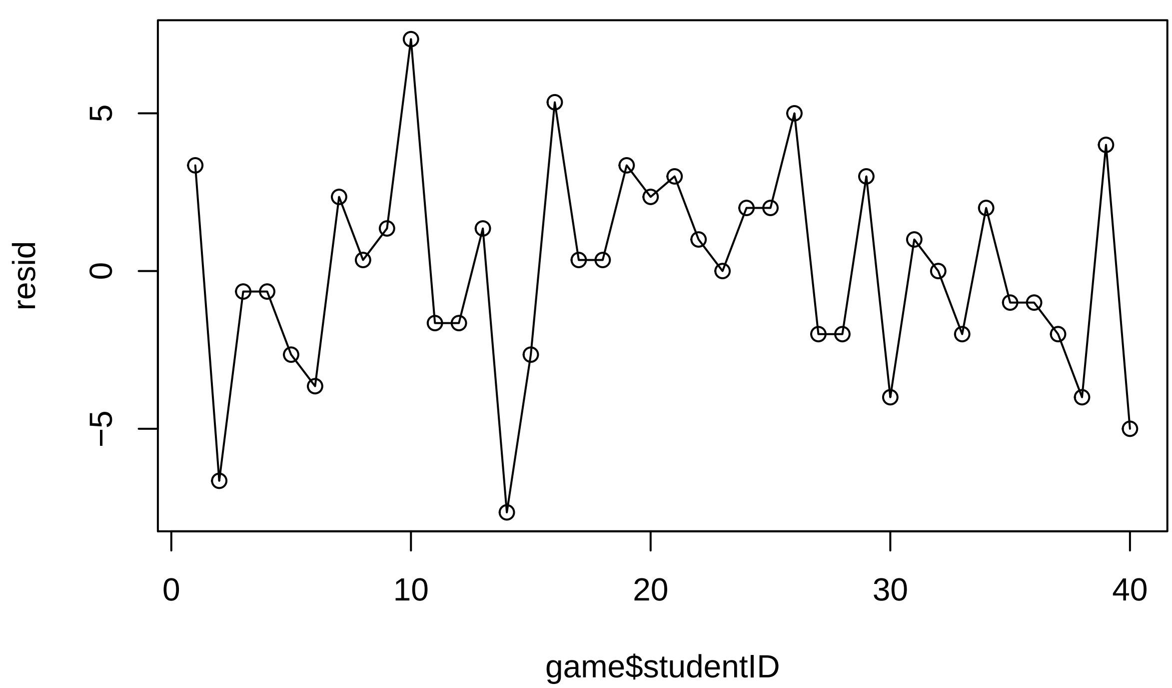
```
## [1] 1.733446
```

Using the informal test, we find that the maximum variance over the minimum variance is 1.733446, which is less than 4, thus we can assume the variances are equal.

4) Plot residuals versus the order of the data to determine if any pattern exist that may indicate the observations are not independent.

```r
resid_left <- left$Time-mean(left$Time)
resid_right <- right$Time-mean(right$Time)

resid <- c(resid_left, resid_right)
plot(resid~game$studentID)
points(resid~game$studentID, type = "l")
```



From the graph above, we find that the residuals against the order do not have any pattern to indicate the observations are not independent.

5) Use a statistical software (t.test()) to conduct a two sample t-test (assuming equal variances) and find

the p-value corresponding to this statistics. Calculate the test statistic by hand and check if its equal with the result from t.test. In addition, use a software to calculate a 95% confidence interval for the difference between the two means $\mu_1 - \mu_2$.

```
t.test(right$Time, left$Time, paired = FALSE, var.equal = TRUE, conf.level = .95, mu = 0, alternative =
```

```
##
##  Two Sample t-test
##
## data:  right$Time and left$Time
## t = -3.5375, df = 38, p-value = 0.001083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.738764 -1.561236
## sample estimates:
## mean of x mean of y
##     35.00     38.65
```

From the t.test(), we find our t value is 3.5375 with a corresponding p value of 0.001083, from which we can reject the null hypothesis. We also find that the 95 percent confidence interval is (-5.738764, -1.561236).

6) Use software instructions and the Game2 data to create indicator variables with $x = 1$ represents the game played with right hand and $x = 0$ represents the left hand game. Develop a regression model using the Time as the response variable and the indicator variable as the explanatory variable.

```
D <- (game$Hand == "Right")*1
model <- lm(game$Time~D)

summary(model)
```

```
##
## Call:
## lm(formula = game$Time ~ D)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.650 -2.000  0.175  2.087  7.350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6500     0.7296  52.975  < 2e-16 ***
## D            -3.6500     1.0318  -3.538  0.00108 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.263 on 38 degrees of freedom
## Multiple R-squared:  0.2477, Adjusted R-squared:  0.2279
## F-statistic: 12.51 on 1 and 38 DF,  p-value: 0.001083
```

Here, we take our boolean array of right = 1 and left = 0 and we fit a linear model using lm(). By using the summary(), we find that our p value for β_1 is less than .05, thus we can conclude $\beta_1 \neq 0$.

7) Use statistical software to calculate the t-statistic and p-value for the hypothesis test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Conduct a 95% confidence interval for β_1. Based on the statistis, can you conclude that the coefficient, β_1, is significantly different from zero?

```
confint(model)
```

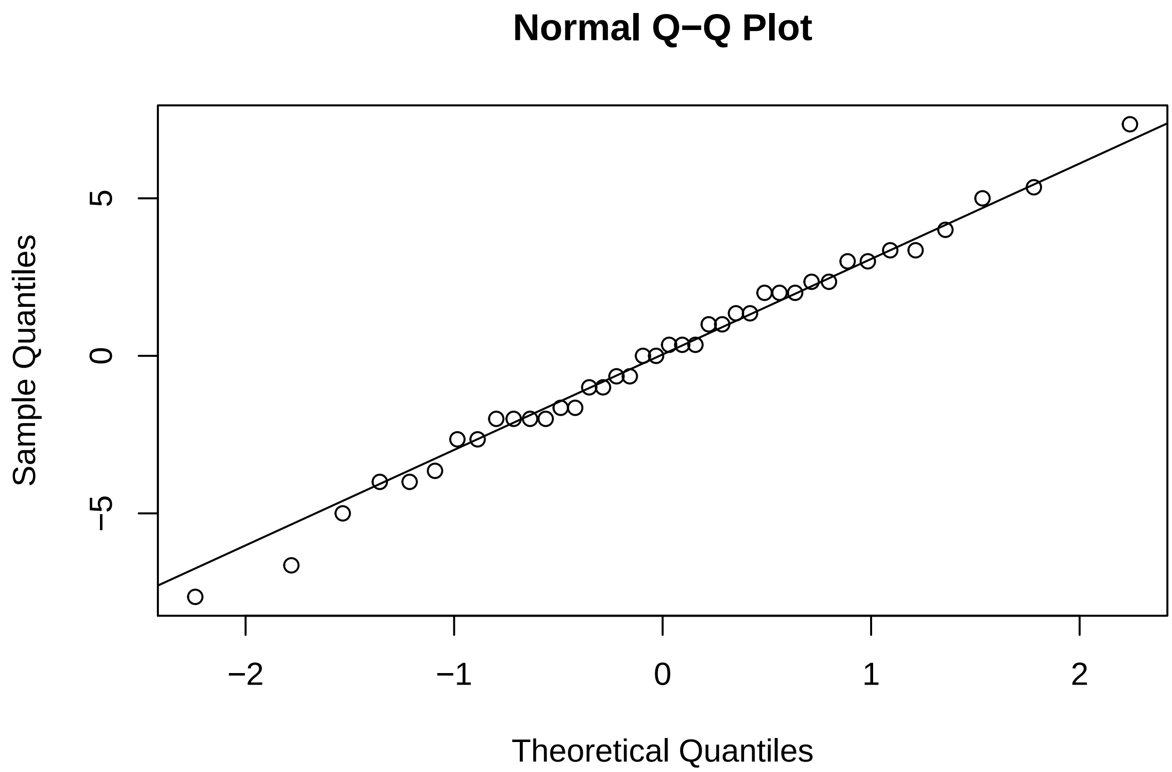
```
##                 2.5 %    97.5 %
## (Intercept) 37.173021 40.126979
## D           -5.738764 -1.561236
```

Based on the confidence interval, we can definitely conclude β_1 is significantly different from zero, as zero lies to the right of the 97.5 percent of our confidence interval.

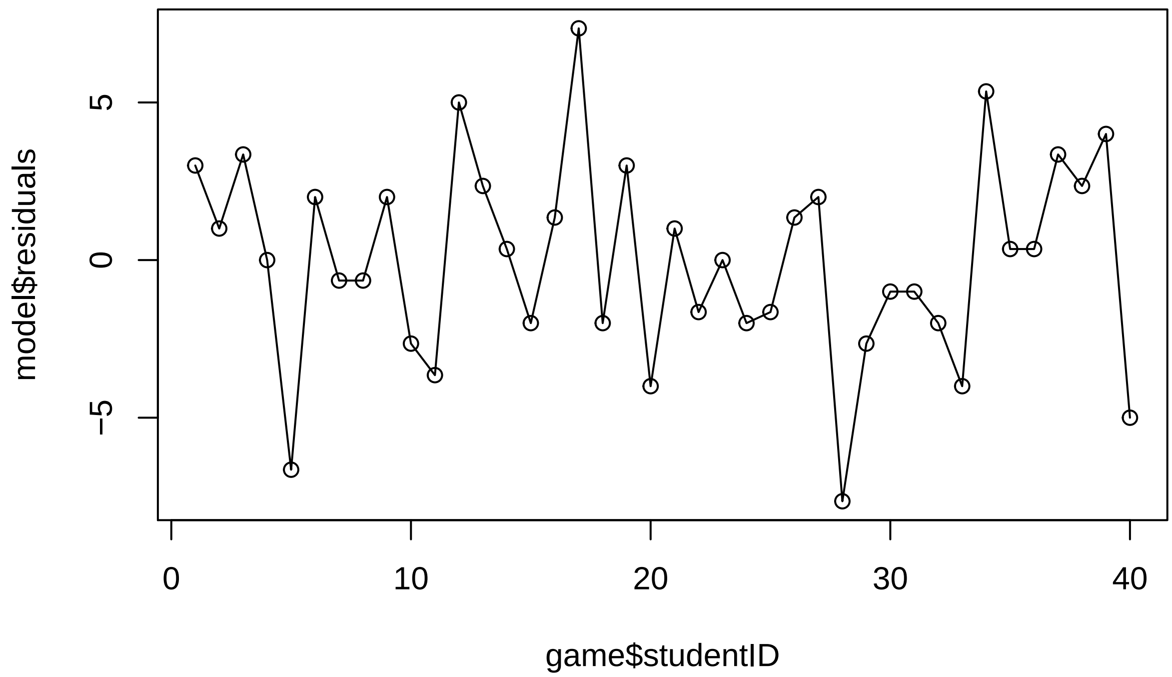
8) Calculate the residuals from the regression line. Plot a histogram of the residuals (or create a normal probability plot of the residuals). Create a residual verus order plot.

```
hist(model$residuals)
```


```
qqnorm(model$residuals)
qqline(model$residuals)
```




```
plot(model$residuals~game$studentID)
points(model$residuals~game$studentID, type = "l")
```



We can find the residuals of our model from model$residuals. From the histogram and the qqnorm plots, we find that the residuals are quite normally distributed. Plotting the residuals against the order, we find that the residuals do not have any kind of pattern, and we conclude our residuals are not only normally distributed, but also iid.

9) Estimate the effect size for the right and left completion time.

```
grand_mean <- mean(game$Time)

right_effect <- mean(right$Time) - grand_mean
left_effect <- mean(left$Time) - grand_mean
right_effect
```

```
## [1] -1.825
```

```
left_effect
```

```
## [1] 1.825
```

From the above, we find the right effect is -1.825 and the left effect is 1.825.

10) Use statistical software to calculate the F-statistics and find the p-value. Use the p-value to draw conclusions from this study.

```
model1 <- aov(game$Time~D)

summary(model1)
```

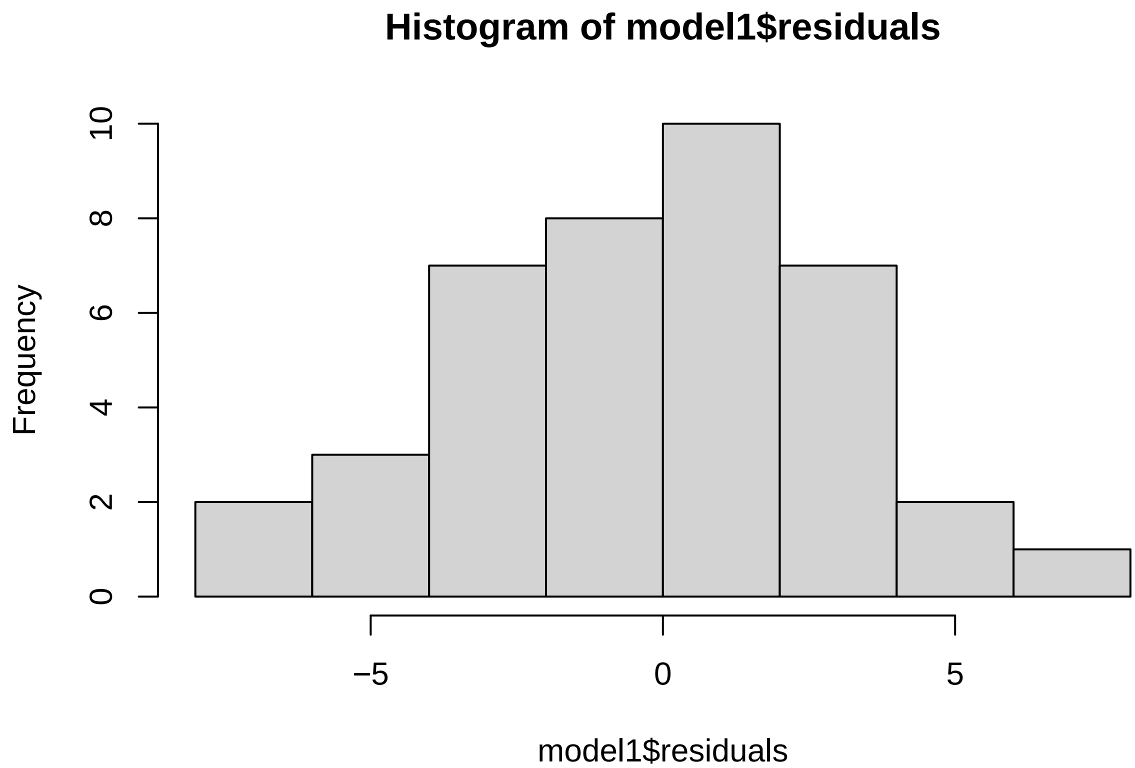
```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## D            1  133.2  133.22   12.51 0.00108 **
```

```
## Residuals   38  404.6   10.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

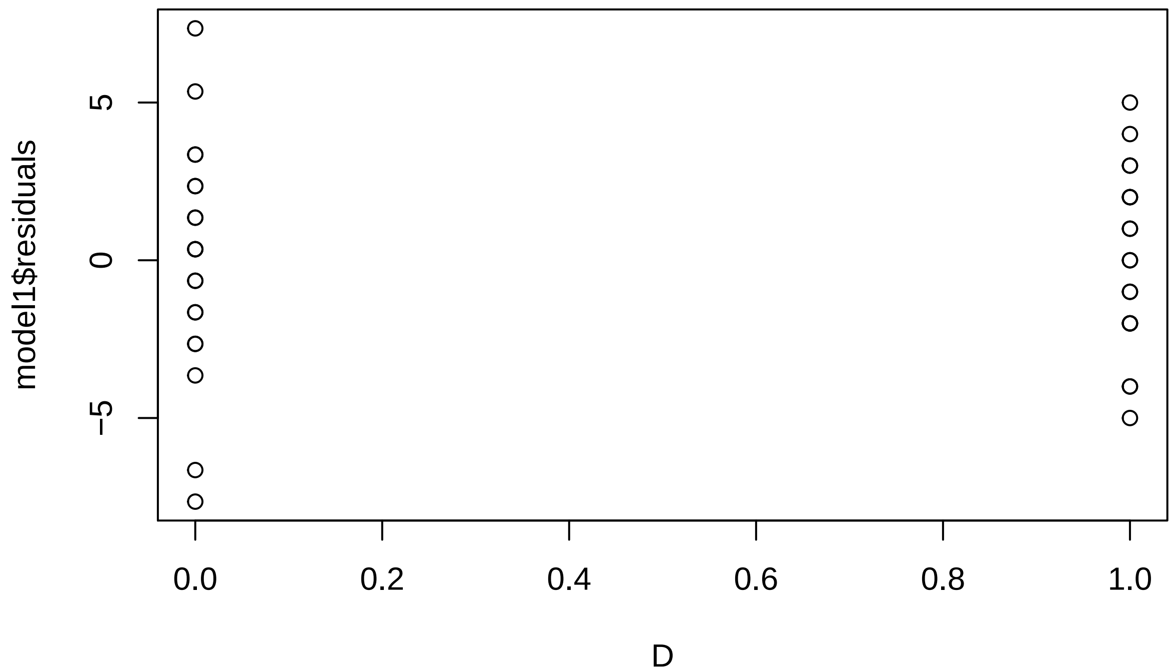
From our ANOVA table, we find that the F value and the p value are the same for the linear regression model.

11) Check the model assmptions by creating a histogram of the residuals, a plot of the residuals versus the hand, and a plot of the residuals versus the order.

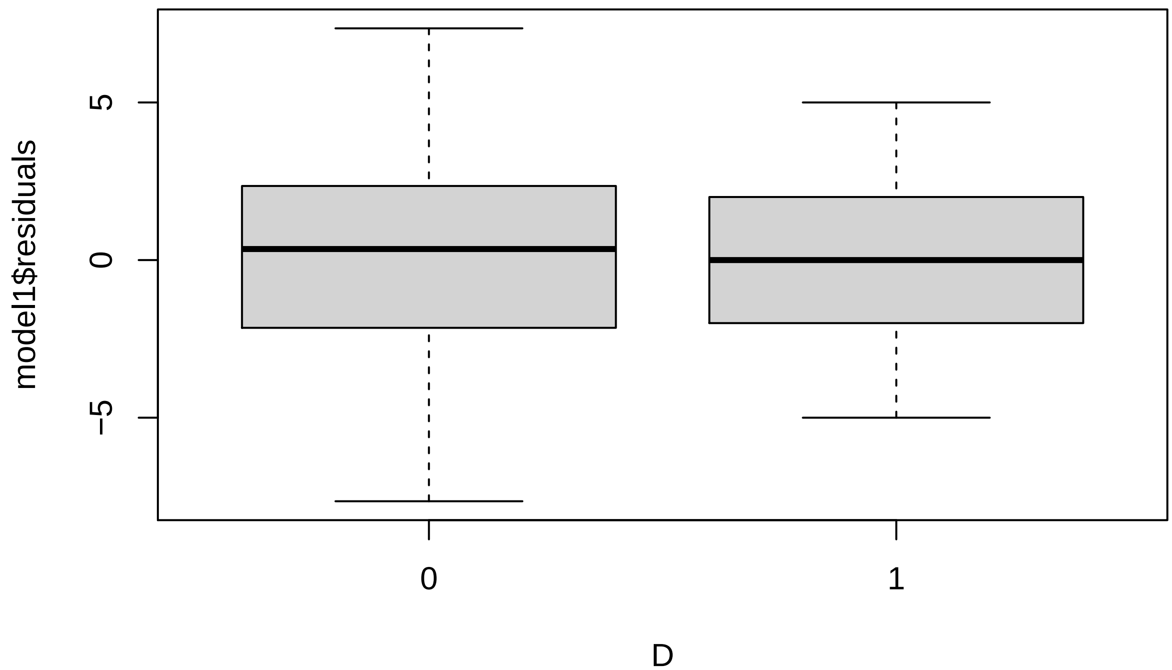
```
hist(model1$residuals)
```



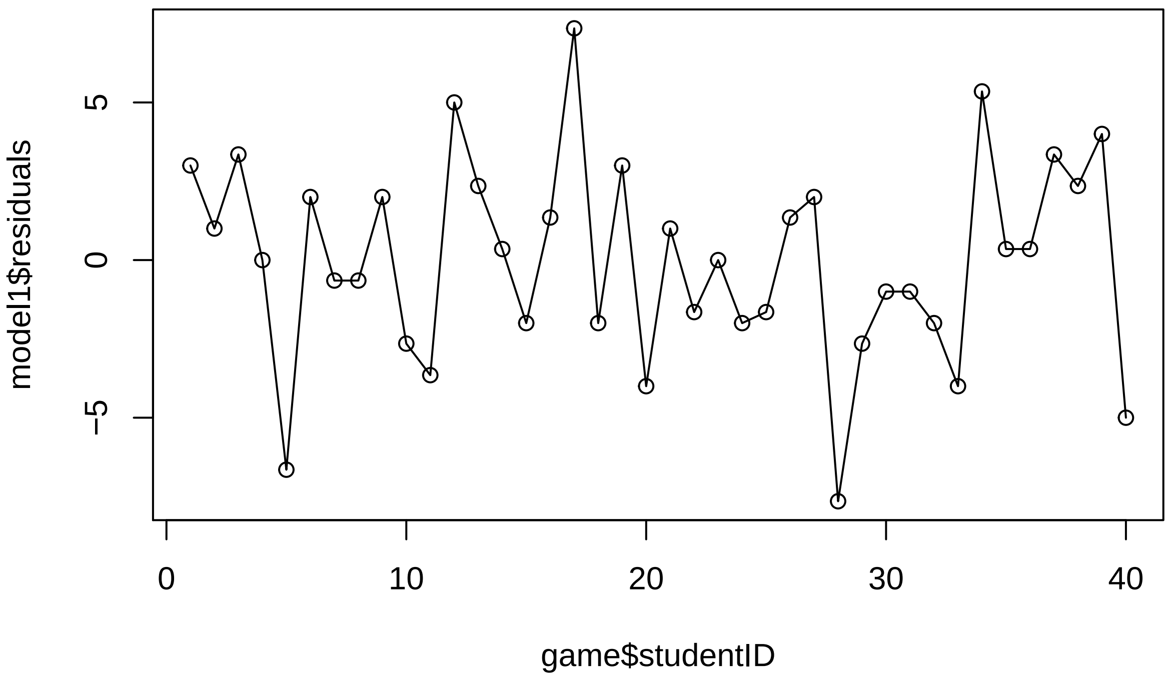

```
plot(model1$residuals~D)
```




```
boxplot(model1$residuals~D)
```




```
plot(model1$residuals~game$studentID, type = "l")
points(model1$residuals~game$studentID)
```



From the histogram above, we find that the residuals seem to be quite normally distributed. From the graph of the residuals against the order, we see there is no pattern in the residuals, and so we conclude the residuals are iid. From the plot of the residuals versus the hand, we find the variance is definitely larger in the left handed times than the right handed times. This is also shown in the boxplot, where we clearly see the $D = 1$ boxplot has a much larger range than the $D = 0$ boxplot.