IMT 573 Lab: Simple Linear Regression

Naga Soundari Balamurugan

November 8th, 2018

{Don't forget to list the full names of your collaborators!}

Collaborators: Jayashree Raman

Instructions:

1. Download the `week7b_lab.Rmd` file from Canvas. Open `week7b_lab.Rmd` in RStudio and supply your solutions to the assignment by editing `week7b_lab.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. If you are using more than just a standard function that you found from another source, please credit the source in the comments.
4. Collaboration on labs is encouraged, but students must turn in an individual assignments. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_Lab7b.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
```

In Class Walk Through

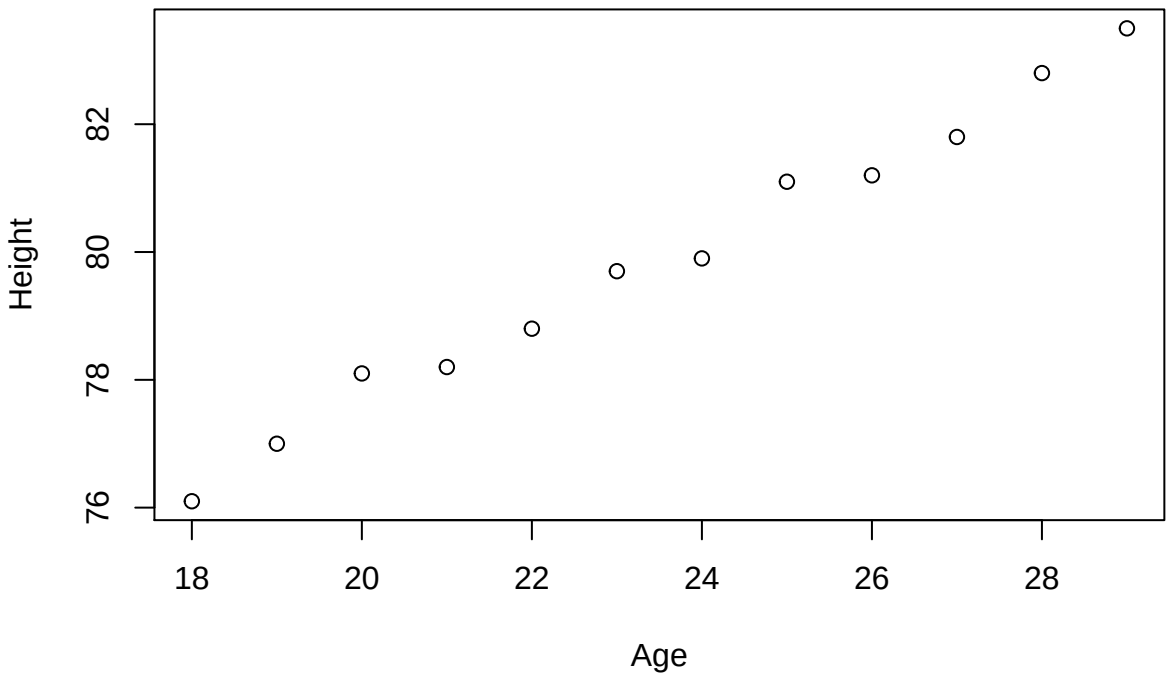
This portion of the lab was adapted from the DataCamp tutorial Linear Regression in R by Eladio Montero Porras, July 18th, 2018.

Download and import the file 'ageheight.csv' from the lab folder in Canvas.

```
# Import csv file
ageheightdata <- read.csv("ageheight.csv")
colnames(ageheightdata) <- c("Age", "Height")
```

Plot the data to make a determination about whether the data appear to have a linear relationship.

```
plot(ageheightdata)
```



Check the correlation between the age variable and the height variable. Note: "There are different methods to perform correlation analysis: Pearson correlation (r), which measures a linear dependence between two variables (x and y). Its also known as a parametric correlation test because it depends to the distribution of the data. It can be used only when x and y are from normal distribution. The plot of y = f(x) is named the linear regression curve. Kendall tau and Spearman rho, which are rank-based correlation coefficients (non-parametric)" (from Correlation Test Between Two Variables in R, STHDA).

```
corTest <- cor(ageheightdata$Age, ageheightdata$Height)
corTest
```

```
## [1] 0.9943661
```

```
corMatrix <- cor.test(ageheightdata$Age, ageheightdata$Height, method = "pearson")
corMatrix
```

```
##
##  Pearson's product-moment correlation
##
## data:  ageheightdata$Age and ageheightdata$Height
## t = 29.665, df = 10, p-value = 4.428e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9793465 0.9984716
## sample estimates:
##       cor
## 0.9943661
```

What does the correlation coefficient tell us?

Age and height has a high positive correlation

Run a linear regression using **height** as the dependent variable (also known as the response variable, predicted variable, or output variable) and **age** as the independent variable (also known as the predictor variable, regressor, explanatory variable, feature, or input variable). Assign the model to a variable. In equation form this looks like:

$Height = \beta_0 + \beta_1 \times Age$

```
lm_ageheight <- lm(Height ~ Age, data = ageheightdata)
summary(lm_ageheight)
```

```
##
## Call:
## lm(formula = Height ~ Age, data = ageheightdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27238 -0.24248 -0.02762  0.16014  0.47238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
## Age           0.6350     0.0214   29.66 4.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.256 on 10 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9876
## F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

Review the results of the model you just ran.

```
lm_ageheight$coefficients
```

```
## (Intercept)         Age
##   64.928322    0.634965
```

```
lm_ageheight$fitted.values
```

```
##        1        2        3        4        5        6        7        8
## 76.35769 76.99266 77.62762 78.26259 78.89755 79.53252 80.16748 80.80245
##        9       10       11       12
## 81.43741 82.07238 82.70734 83.34231
```

We now have β_0 and β_1 parameters for our model. Which parameter is considered the intercept and which is considered the slope? Add these in to the equation from above ($height = \beta_0 + \beta_1 \times age$):

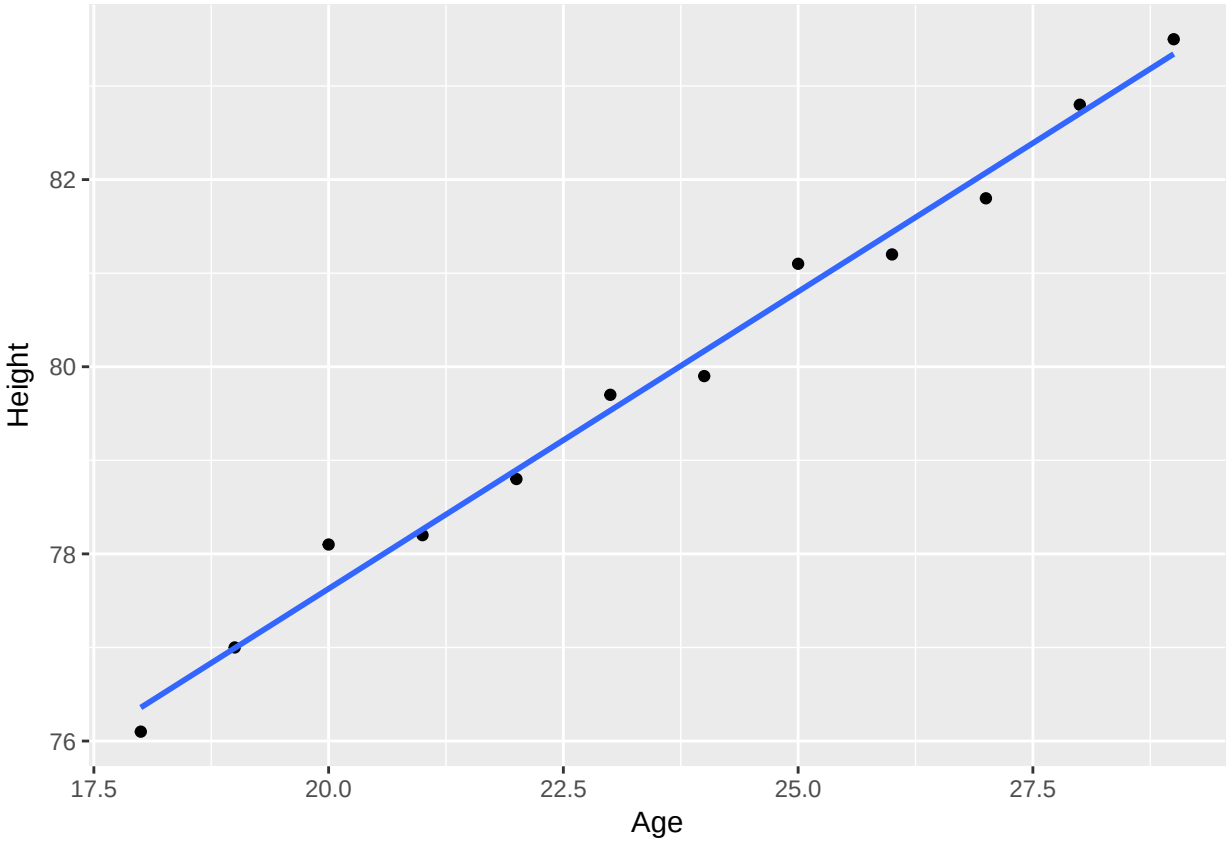
> β_0 is the intercept and β_1 is the slope. The value of intercept is 64.928322 and the slope is 0.634965.

The equation looks like, $Height = 64.928322 + (0.634965 * Age)$

Plot the data with the linear regression line.

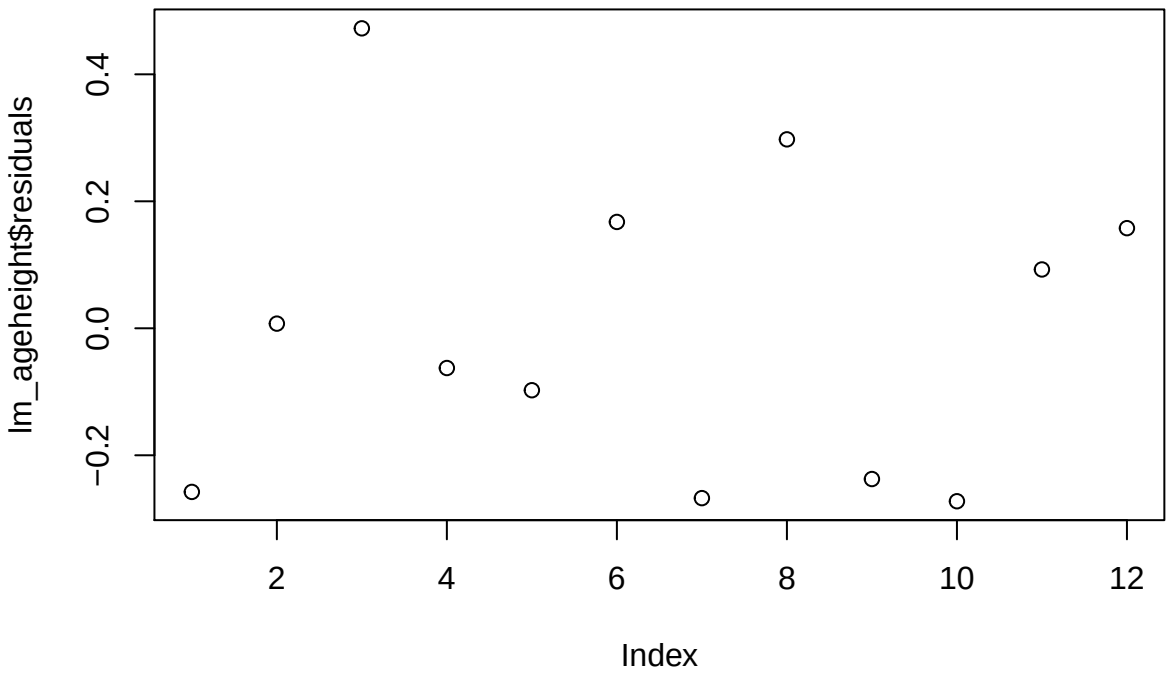
```
# with(ageheightdata, plot(Age, Height))
#   abline(lm_ageheight, col = 'green')

ggplot(ageheightdata, aes(x = Age, y = Height)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE)
```

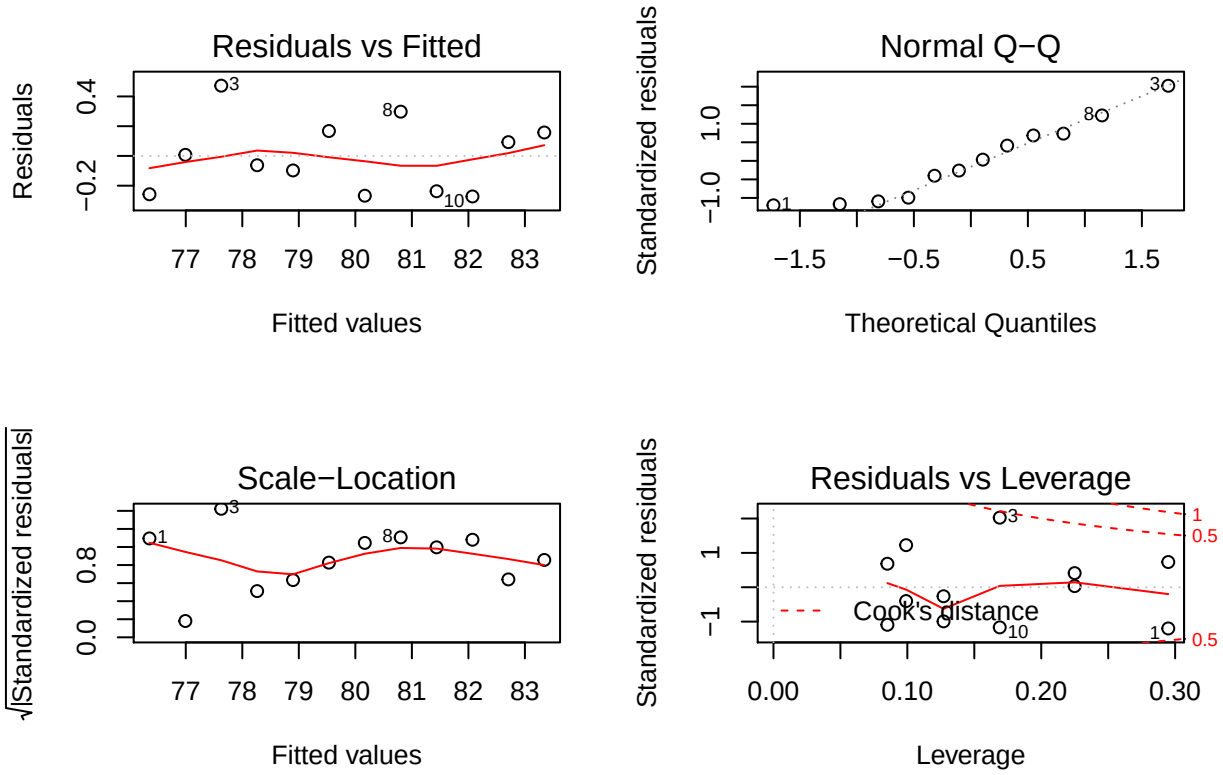


Let's check the residuals plot to see if the linear regression assumptions about residuals have been met. (Useful reference: https://uc-r.github.io/linear_regression and https://data.library.virginia.edu/diagnostic-plots/)

```
plot(lm_ageheight$residuals)
```




```
par(mfrow = c(2,2))
plot(lm_ageheight)
```




```
#Normal Q-Q: Tells about "Are the residuals normally distributed"
#Sclae-Location plot: Tells if we have equal variance.
#Residuals vs Leverage: Tells If there are influential datapoints - if the datapoint has influence to s
```

Based on our model, calculate the predicted height $\hat{height}$ of a child who is 1 month old, 26 months old, and 60 months old. Do not use a prediction function in R, do these calculations with simple math in R or by hand and report your results below:

```
Height_1Mon = 64.928322 + (0.634965 * 1)
Height_1Mon
```

```
## [1] 65.56329
```

```
Height_26Mon = 64.928322 + (0.634965 * 26)
Height_26Mon
```

```
## [1] 81.43741
```

```
Height_60Mon = 64.928322 + (0.634965 * 60)
Height_60Mon
```

```
## [1] 103.0262
```

```
# predict the height in centimeters for the following dataframe of age values using
# a the prediction function in r (look at predict())
testSet <- data.frame(Age=c(0, 21.25, 30.5, 120, 360, 1200))

predValues <- predict(lm_ageheight, newdata = testSet)
predValues
```

```
##         1         2         3         4         5         6
##  64.92832  78.42133  84.29476 141.12413 293.51573 826.88636
```

Are your prediction results problematic? Where did we go wrong?

> Yes, the height of the person whose age is extremely high or low is not real. It is because we have built a model whose data range is 18 - 29 months. Thus making the model predict height of the person whose age is above or beyond the range could not be right.

In Class Lab

Sports Statistics: Predicting Runs Scored in Baseball

Baseball is a played between two teams who take turns batting and fielding. A run is scored when a player advances around the bases and returns to home plate. The data we will use today is from all 30 Major League Baseball teams from the 2011 season. This data set is useful for examining the relationships between wins, runs scored in a season, and a number of other player statistics.

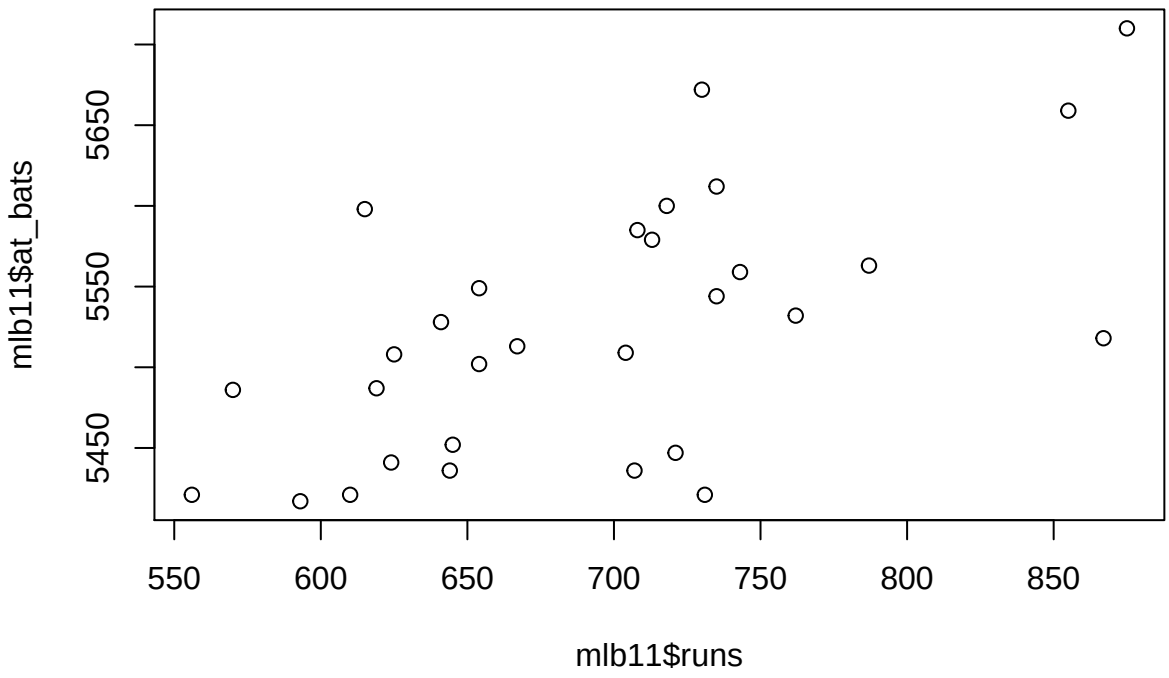
\marginnote{Note: More info on the data can be found here: https://www.openintro.org/stat/data/mlb11.php

```
# Download and load data
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")
load("mlb11.RData")
```

Use the baseball data to answer the following questions:

- Plot the relationship between runs and at_bats. Does the relationship look linear? Describe the relationship between these two variables.

```
plot(mlb11$runs, mlb11$at_bats)
```



The relationship is slightly linear. The runs increases with the increase in at_bats.

- If you knew a teams at bats, would you be comfortable using a linear model to predict the number of runs?

  Yes the number of runs could be predicted with model.

- If the relationship looks linear, quantify the strength of the relationship with the correlation coefficient. Discuss what you find.

```
cor_runs_atbats <- cor.test(mlb11$runs, mlb11$at_bats)
cor_runs_atbats
```

```
##
##  Pearson's product-moment correlation
##
## data:  mlb11$runs and mlb11$at_bats
## t = 4.0801, df = 28, p-value = 0.0003388
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3209675 0.7958231
## sample estimates:
##      cor
## 0.610627
```

These two variables are positivey correlated. Though the correlation is not so high, its has above moderate correlation(0.6106).

- Use the `lm()` function to fit a simple linear model for runs as a function of at bats. Write down the

formula for the model, filling in estimated coefficient values.

```
lm_runs <- lm(runs ~ at_bats, data = mlb11)
summary(lm_runs)
```

```
##
## Call:
## lm(formula = runs ~ at_bats, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.58  -47.05  -16.59   54.40  176.87
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2789.2429   853.6957  -3.267 0.002871 **
## at_bats         0.6305     0.1545   4.080 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.47 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

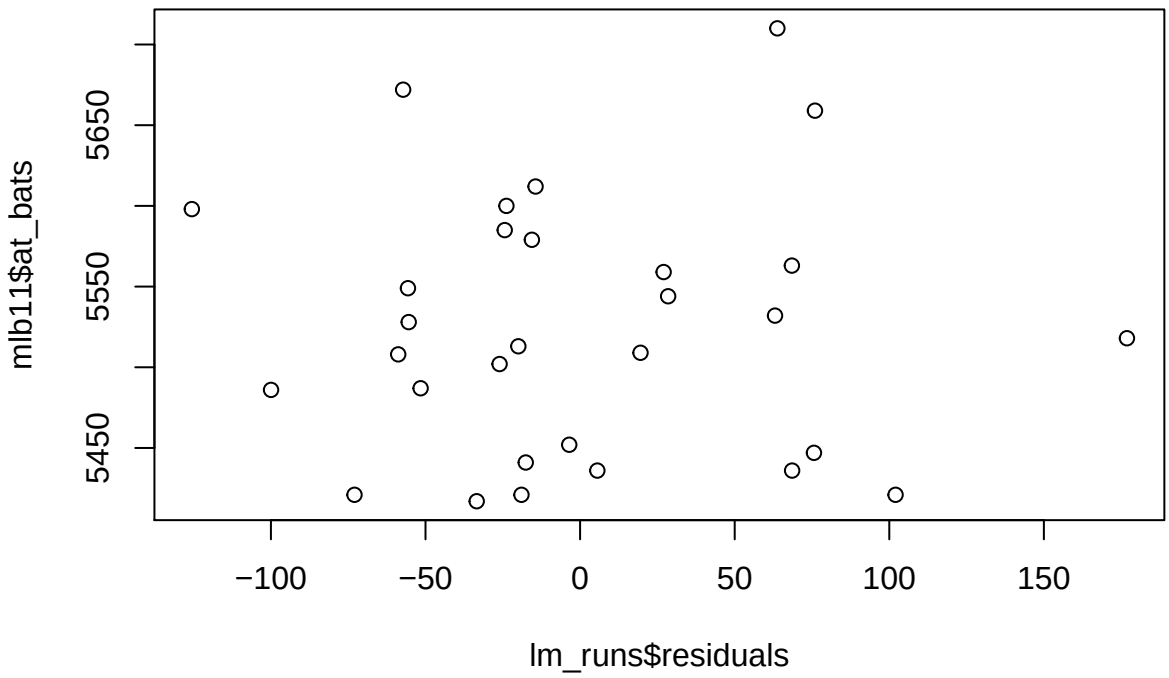
The formula that could be used to calculate the runs by at_bats is $Runs = (-2789.2429) + (0.6305 * at\_bats)$

- Describe in words the interpretation of β_1.

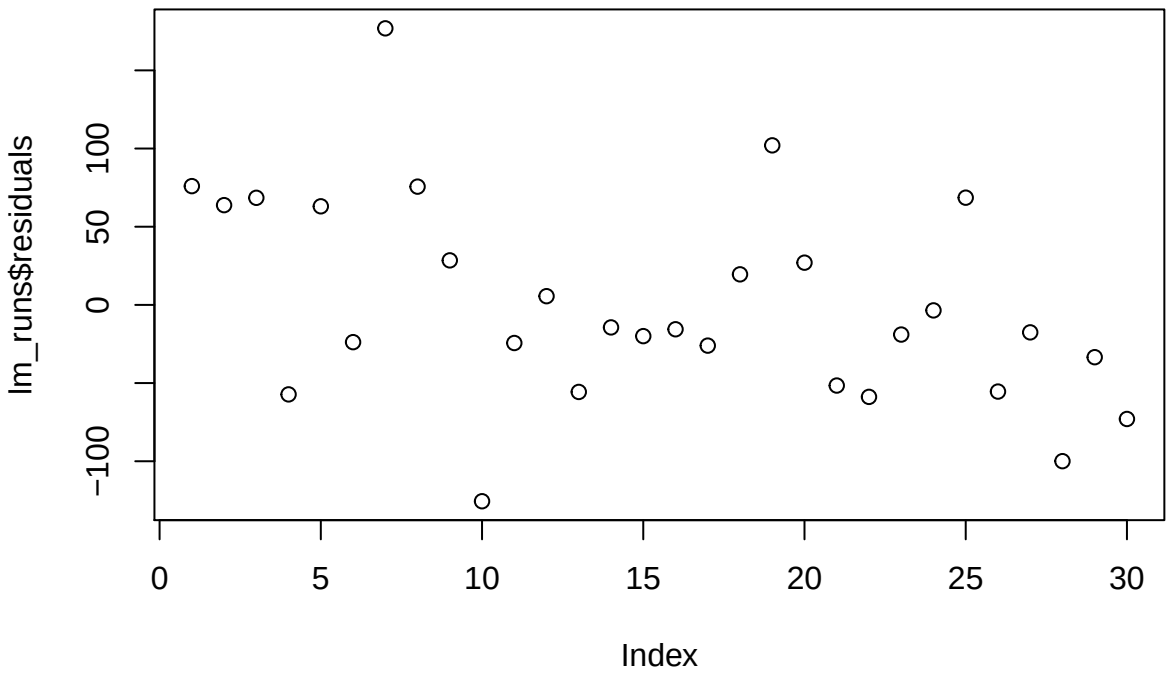
  The no of runs increases/decreases by 0.6305 times the increase/decrease of at_bats respectively. In addition to the β_1 variable, the value of β_0 should be added to it.

- Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot?

```
plot(lm_runs$residuals, mlb11$at_bats)
```




```
plot(lm_runs$residuals)
```



- Comment on the fit of the model.

  The residual plot looks pretty normal as expected. The data points are spread evenly across the 0 except a single outlier which is at 150. Thus this model looks like a good fit for the data.