# ACTIVITY-08

***Formulate the Hypothesis function for Linear Regression and Investigate the computational analysis of the linear regression model to estimate the coefficients for any real-world application.***

**Hypothesis function for Linear Regression:**

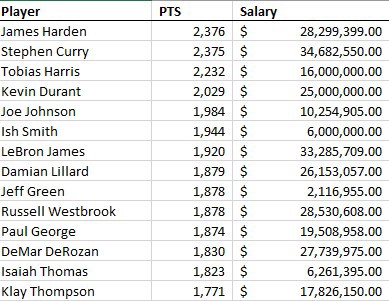
**hθ = θ1 + θ2x**

* Our hypothesis function is exactly the same as the equation of a line using the slope and y-intercept. **y=mx+b**
* θ1: intercept
* θ2: coefficient of x
* Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

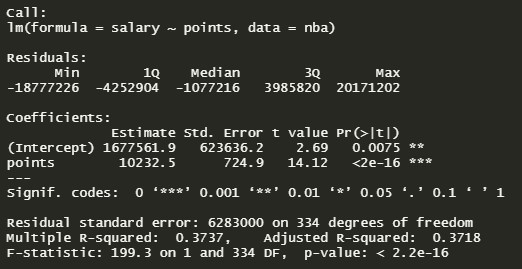
**Linear regression Applications:**

* In linear regression in ML we assume that y and x are related with the following equation: **y = wx+ε** Given an input x we would like to compute an output y, where w is a parameter and ε represents measurement error or other noise.
* There are different techniques to estimate the parameters of a model. One of the most popular is the **Ordinary Least Squares**.
* The premise of the Ordinary Least Squares method is to minimize the sum of the squares of the **residuals** of the model. Which is the difference between the predicted values and the actual values in the dataset.
* This way the model is calculating the best parameters, so that each point in the regression line is as *close* as possible to the dataset.

using a [dataset](https://www.kaggle.com/koki25ando/salary) from the National Basketball Association (NBA). This dataset includes salary information and points scored during the season for each player in the 2017–2018 season. We’ll be investigating the relationship between points scored in a season and the salary of a player.



by running a simple regression model with salary as our dependent variable and points as our independent variable. The output of this regression model is below:



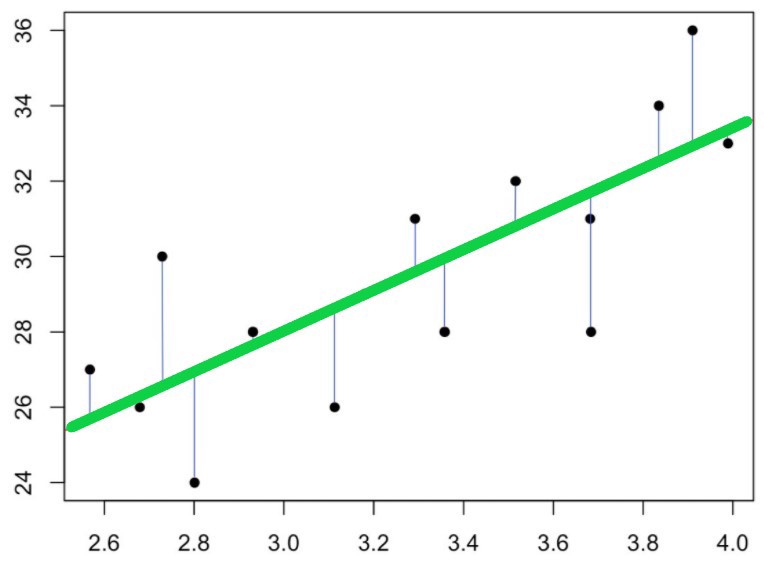
The call section shows us the formula that R used to fit the regression model. Salary is our dependent variable and we are using points as a predictor (independent variable) from the NBA dataset.

The residuals are the difference between the actual values and the predicted values. We can generate these same values by taking the actual values of salary and subtracting it from the predicted values of the model:

summary(nba$salary - model$fitted.values)



To understand what the coefficients are, we need to go back to what we are actually trying to do when we build a linear model. We are looking to build a generalized model in the form of y=mx+b, where b is the intercept and m is the slope of the line. Because we often don’t have enough information or data to know the exact equation that exists in the wild, we have to build this equation by generating estimates for both the slope and the intercept. These estimates are most often generated through the ordinary least squares method.



It is from this line above that we obtain our coefficients. Where the line meets the y-axis is our intercept (b) and the slope of the line is our m. Using the understanding we’ve gained so far, and the estimates for the coefficients provided in the output above, we can now build out the equation for our model. We’ll substitute points for m and (Intercept) for b:

y=$10,232.50(x) + $1,677,561.90

 if an NBA player scored zero points during a season, that player would make $1,677,561.90 on average. Then, for each additional point they scored during the season, they would make $10,232.50.

**Coefficients — Std. Error**

The standard error of the coefficient is an estimate of the standard deviation of the coefficient. In effect, it is telling us how much uncertainty there is with our coefficient. The standard error is often used to create confidence intervals. For example we can make a 95% confidence interval around our slope, points:

$10,232.50 ± 1.96($724.90) =($8,811.70,$11,653.30)

Looking at the confidence interval, we can say we are 95% confident that the actual slope is between $8,811.70 and $11,653.30.