Project Report On Regression Techniques And Parameter Estimation

Submitted By:

Ankit Desai
Mehul Katara(1421005)
Parayagraj Shah(1421009)
Sahil Desai(1421014)
Saurabh Chauhan(1421015)

Guided By:

Prof. Mehul Raval
Dhruv Gupta

Contents

List of Figures

Introduction

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables de- noted X. The case of one explanatory variable is called simple linear regression or univariate linear regression.

The method of **least squares** is a standard approach to the approximate solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. Least squares" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation.

Random sample consensus (RanSaC) is an iterative method to estimate pa- rameters of a mathematical model from a set of observed data which contains outliers. It is a non-deterministic algorithm in the sense that it produces a rea- sonable result only with a certain probability, with this probability increasing as more iterations are allowed. The algorithm was first published by **Fischler and Bolles** at SRI International in 1981.

Gradient Descent method is a way to find a local minimum of a function. The way it works is we start with an initial guess of the solution and we take the gradient of the function at that point. We step the solution in the negative direction of the gradient and we repeat the process. The algorithm will eventually converge where the gradient is zero (which correspond to a local minimum).

In statistics, **maximum-likelihood estimation (MLE)** is a method of estimat- ing the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

1 Overview of Mathematical Model

1.1 Linear regression

- Linear Regression is implemented using Least Square Estimation.
- Using Least Square , Slope is given by
  - Slope=$(\sum((x-\bar{x})*(y-\bar{y})))/(\sum(x-\bar{x})^2)$
  - Intercept=$(\bar{y}-(Slope*\bar{x}))$(Because Regression Line should passed through $\bar{x}$ and $\bar{y}$)
- $y = Intercept + (Slope * x)$
- Once Intercept and Slope are known to us , we can find expected value of y corresponding value of x.

1.2 RanSaC

- Select randomly the minimum number of points required to determine the model parameters.
- Solve for the parameters of the model.
- Determine how many points from the set of all points fit with a predefined tolerance ϵ.
- If the fraction of the number of inliers over the total number points in the set exceeds a predefined threshold τ,re-estimate the model parameters using all the identified inliers and terminate.
- Otherwise,repeat steps 1 through 4(maximum of Ntimes).

1.3 Gradient Descent

- In Gradient Descent, We have to minimize the cost function which is given as below,

$$J(\theta_0,\theta_1) \quad = \quad argmin(\theta_0,\theta_1)\sum_{i=1}^{m}(h_\theta(x^{(i)})-(y^{(i)}))^2$$

Repeat until convergence (for every j)

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m}\left(y^{(i)} - h_\theta x^{(i)}\right)(x_j)^{(i)}$$

1.4 Maximum-likelihood Estimation

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model.
- Here assumption is that it follows Gaussian Distribution so our equation is as below

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{m}(1/\sqrt{2\pi}\sigma)\exp\left(-(y^{(i)}-\theta^T x^{(i)})^2/(2*\sigma^2)\right)\\
\textit{l(}\theta\textit{)} &= \log L(\theta)\\
&= \log\prod_{i=1}^{m}(1/\sqrt{2\pi}\sigma)\exp\left(-(y^{(i)}-\theta^T x^{(i)})^2/(2*\sigma^2)\right)\\
&= \sum_{i=1}^{m}\log(1/\sqrt{2\pi}\sigma)\exp\left(-(y^{(i)}-\theta^T x^{(i)})^2/(2*\sigma^2)\right)\\
&= m\log(1/\sqrt{2\pi}\sigma)-(1/\sigma^2).(1/2)\sum_{i=1}^{m}\left(-(y^{(i)}-\theta^T x^{(i)})^2\right)
\end{aligned}$$

- Hence, maximizing $l(\theta)$ gives the same answer as minimizing

$$= (1/2)\sum_{i=1}^{m}\left(-(y^{(i)}-\theta^T x^{(i)})^2\right)$$

2 Problem Statement:1

2.1 Introduction

- A snack is a portion of food often smaller than a regular meal, generally eaten between meals. Snacks come in a variety of forms including packaged and processed foods and items made from fresh ingredients at home.

- Collect the data from the local supermarket about Fat Content and Calories in Snack Foods (at least 25 products). Perform linear regression over the data collected using Least Square Estimation and RanSaC

- Sample Data are shown Below:

Sr. No.	Item Name	Calories(K cal)	Fat(gm)
1	MAGGIE	503	23.3
2	MAGGIE PASTA	402	14.4
3	BINGO CHIPS	529.97	29.55
4	PARLE CHIPS	501.72	22.58
5	UNCLE CHIPS	540.58	32.36

(K cal per 100 gm)

2.2 Sample Calculation using Least Square & RanSaC

- Sample calculations are as Below:

Calorie	Fat	Expected Fat(using Least Square)	Expected Fat (Using RanSaC)
503	23.3	28.66	23.55
402	14.4	18.06	16.48
529.97	29.55	31.49	25.4379

- Data is collected from Local Market of Ahmedabad.
- Results of Least Square Estimation and RanSaC are Shown Below (Output of Mathematical Models are Shown Below):

	Least Square estimation	RanSac
Intercept	-24.14286	-11.660059
Slop	.10499	0.070671
Residual standard error	7.067	4.806
Multiple R-squared	.7064	.7076
Correlation Coefficient	.8404692	0.7017397

2.3 Results in form of Graphs

- Here in the above Graph Red Line represent Expected Fat using Least Square Estimation and Blue Line represent Expected Fat using RanSaC implementation.

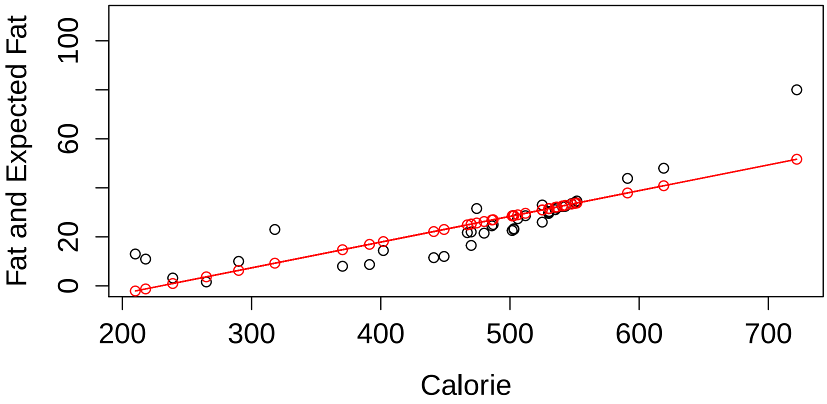

Figure 1: Calorie vs Fat/Expected Fat

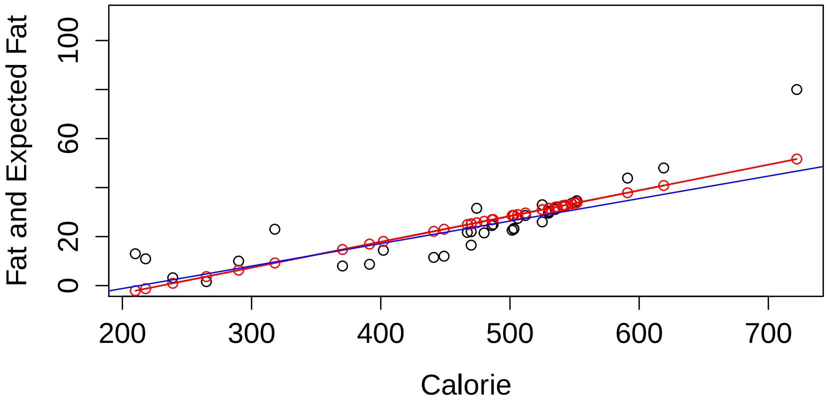

Figure 2: Graph of Least Square and RanSaC output

3 Problem Statement:5

3.1 Introduction

- Shane Keith Warne (born 13 September 1969) is an Australian former international cricketer widely regarded as one of the best bowlers in the history of the game. He was named one of the Wisden Cricketers of the Year in the 1994 Wisden Cricketers' Almanack. He was the Wisden Leading Cricketer in the World 1997 (Notional Winner). He was named Wisden Leading Cricketer in the World for the year 2004 in 2005 Wisden Cricketers' Almanack.
- Dig out the data about ODI career of Shane Warne. Collect information about the Runs given and Wickets taken by Shane Warne in a cricket match (at least 75 ODI matches). Perform linear regression over the data collected using Least Square Estimation and RanSaC.
- Sample Data are shown Below:

Match. No.	Run Given	Wicket Taken
1	40	2
2	43	1
3	25	4
4	19	4
5	27	2

3.2 Sample Calculation using Least Square & RanSaC

- Sample calculations are as Below:

Run	Wicket	Exp Wicket (using Least Square)	Exp Wicket (Using RanSaC)
40	2	28.66	23.55
43	1	18.06	16.48
25	4	31.49	25.4379

- Data is collected from website of howzthat.
- Results of Least Square Estimation and RanSaC are Shown Below
(Output of Mathematical Models are Shown Below):

	Least Square estimation	RanSac
Intercept	2.9758	2.9758
Slop	-.03477	-.03477
Residual standard error	1.206	1.206
Multiple R-squared	.0891	.0891
Correlation Coefficient	-.2984	-.2984

3.3 Results in form of Graphs

- Here in Below Graph Red Line represent Expected Wicket Taken by S.Warne using Least Square Estimation and Blue Line represent Expected Wicket Taken by S.Warne using RanSaC implementation.

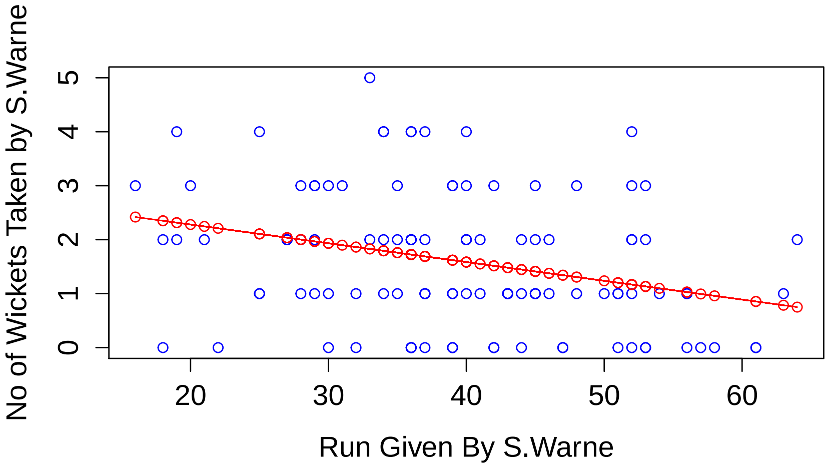

Figure 3: Run Given by S.Warne vs No of Wicket Taken by S.Warne

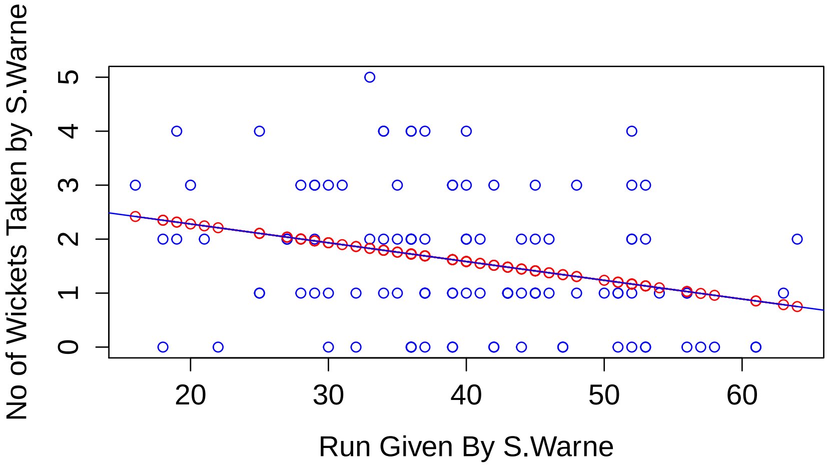

Figure 4: Graph of Least Square and RanSaC Implementation

4 Problem Statement:9(Bonus Problem)

4.1 Introduction

- FIA Formula One World Championship (also Formula One, Formula 1, and F1) is the highest class of single-seat auto racing that is sanctioned by the Fdration Internationale de l'Automobile (FIA). The formula",designated in the name, refers to a set of rules with which all participants' cars must comply. The F1 season consists of a series of races, known as Grands Prix (from French, originally meaning great prizes), held through- out the world on purpose-built circuits and public roads. The results of each race are evaluated using a points system to determine two annual World Championships, one for the drivers and one for the constructors. The racing drivers, constructor teams, track officials, organisers, and circuits are required to be holders of valid Super Licences, the highest class of racing licence issued by the FIA.

- Go through the points table of F1 for year 2012 & 2013 and collect the data of F1 drivers (at least 15 drivers). Plot the average points earned by each driver in a GP against his salary. Perform linear regression over the data collected using Least Square Estimation and RanSaC.

- Sample Data are shown Below:

Driver Name.	Points(2012)	Points(2012)	Salary
Sebastian Vettel	397	281	31.7
Fernando Alonso	242	278	31.70
Kimi Rikknen	183	207	31.7
Lewis Hamilton	189	190	28.6
Mark Webber	199	179	16

Salary are in Million Dollar

4.2 Sample Calculation using Least Square & RanSaC

- Sample calculations are as Below:

Avg Point	Salary	Exp Salary (using Least Square)	Exp Salary (Using RanSaC)
339	31.7	38.01	42.21
260	31.7	29.27	32.23
195	31.7	22.092	24.024

- F1 Players points are collected from Official Website of Formula 1 and Salary is from Business Book 2013.

- Results of Least Square Estimation and RanSaC are Shown Below
(Output of Mathematical Models are Shown Below):

	Least Square estimation	RanSac
Intercept	.53126	-.60231
Slop	.11057	.12629
Residual standard error	5.442	5.195
Multiple R-squared	.8072	.8039
Correlation Coefficient	.8984	.8984

4.3 Results in form of Graphs

- Here in Below Graph Red Line represent Expected Wicket Taken by S.Warne using Least Square Estimation and Blue Line represent Expected Wicket Taken by S.Warne using RanSaC implementation

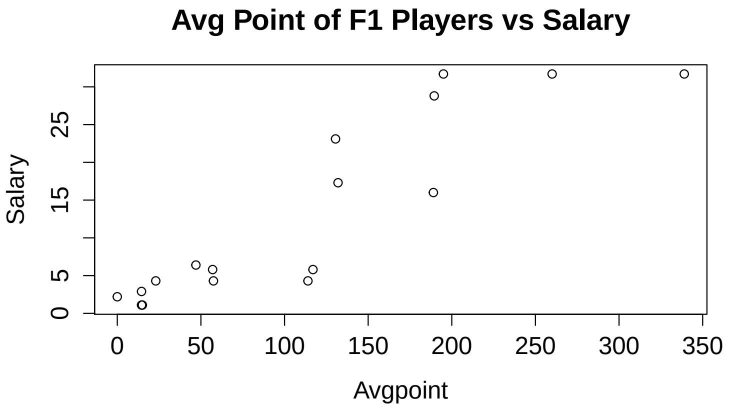

Figure 5: Average points earned by each driver against his Salary.

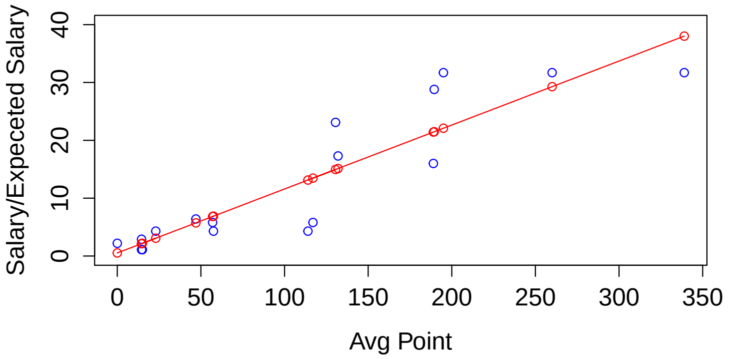

Figure 6: Average points of each driver vs Salary.

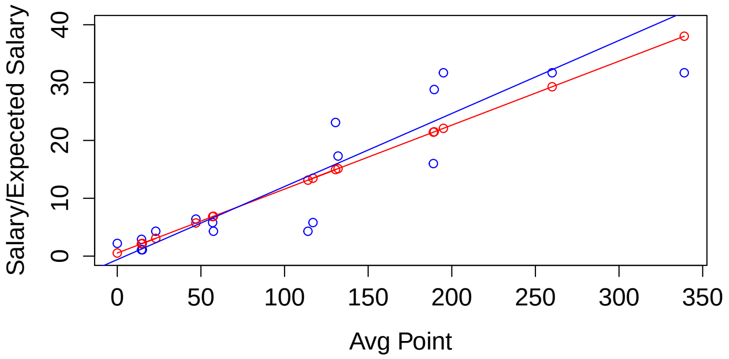

Figure 7: Graph of Least Square and RanSaC Implementation

5 Problem Statement:1(Using Cubic Fit)

5.1 Introduction

- A snack is a portion of food often smaller than a regular meal, generally eaten between meals. Snacks come in a variety of forms including packaged and processed foods and items made from fresh ingredients at home.

- Collect the data from the local supermarket about Fat Content and Calories in Snack Foods (at least 25 products). Draw Cubic Fit over collected data.

- Sample Data are shown Below:

Sr. No.	Item Name	Calories(K cal)	Fat(gm)
1	MAGGIE	503	23.3
2	MAGGIE PASTA	402	14.4
3	BINGO CHIPS	529.97	29.55
4	PARLE CHIPS	501.72	22.58
5	UNCLE CHIPS	540.58	32.36

(K cal per 100 gm)

5.2 Sample Calculation using Cubic Fit

- Sample calculations are as Below:

Calorie	Fat	Expected Fat(using Cubic Fit)
503	23.3	25.58
402	14.4	13.29
529.97	29.55	30.17

- Data is collected from Local Market of Ahmedabad.

- Result of Cubic Fit is Shown Below
(Output of Mathematical Models are Shown Below):

	Cubic Fit
Intercept	3.333e+01
Slop	-1.713e-01
Slop	2.713e-04
Slop	7.680e-08
Residual standard error	4.083
Multiple R-squared	0.92
Correlation Coefficient	0.8404692

5.3 Results in form of Graphs

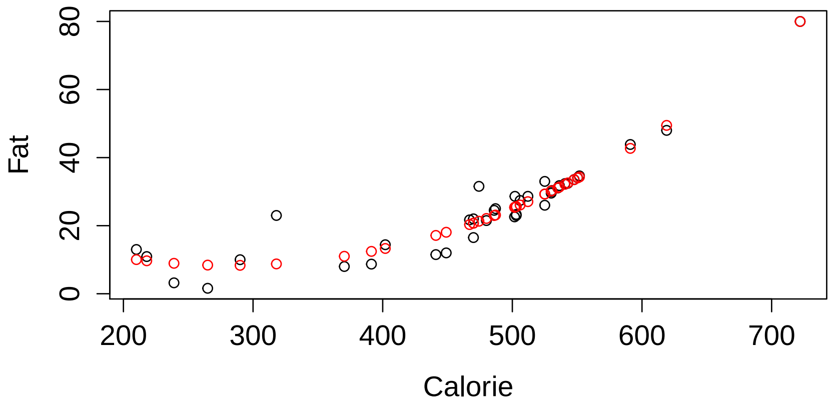

Figure 8: Graph of Cubic Fit

- Here in Above Graph Red Dot represents Expected Fat using Cubic Fit and Black Dot represents Actual value of Fat.

6 Problem Statement:1(Using Gradient Descent)

6.1 Introduction

- A snack is a portion of food often smaller than a regular meal, generally eaten between meals. Snacks come in a variety of forms including packaged and processed foods and items made from fresh ingredients at home.
- Collect the data from the local supermarket about Fat Content and Calories in Snack Foods (at least 25 products). Apply Gradient Descent over collected data.
- Sample Data are shown Below:

Sr. No.	Item Name	Calories(K cal)	Fat(gm)
1	MAGGIE	503	23.3
2	MAGGIE PASTA	402	14.4
3	BINGO CHIPS	529.97	29.55
4	PARLE CHIPS	501.72	22.58
5	UNCLE CHIPS	540.58	32.36

(K cal per 100 gm)

6.2 Sample Calculation using Gradient Descent

- Sample calculations are as Below:

Calorie	Fat	Expected Fat(using Gradient Descent)
503	23.3	25.58
402	14.4	13.29
529.97	29.55	30.17

- Result of Gradient Descent is Shown Below
(Output of Mathematical Models are Shown Below):

	Cubic Fit
Intercept	3.333e+01
Slop	-1.713e-01
Slop	2.713e-04
Slop	7.680e-08
Residual standard error	4.083
Multiple R-squared	0.92
Correlation Coefficient	0.8404692

6.3 Results in form of Graphs

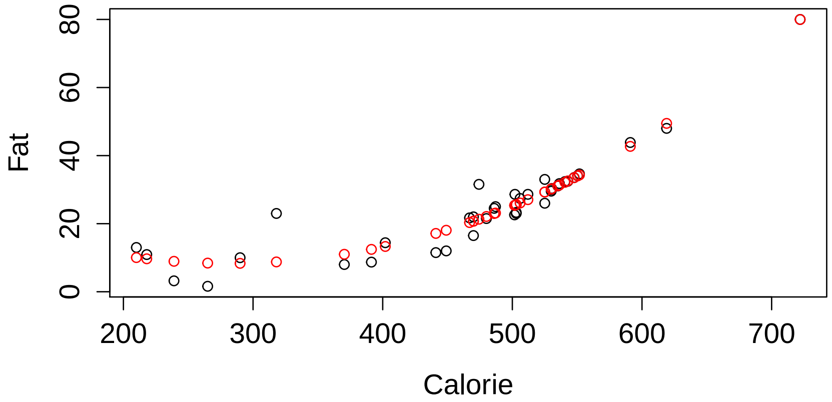

Figure 9: Graph of Gradient Descent

- Here in Above Graph Red Dot represents Expected Fat using Gradient Descent and Black Dot represents Actual value of Fat.

7 Conclusion

- Conclusion through Table is shown below.

	Least Square estimation	RanSac	Cubic Fit	Gradient Descent
Intercept	-24.14286	-11.660059	3.333e+01	3.333e+01
Slop	.10499	0.070671	-1.713e-01	-1.713e-01
Slop1	—	—	2.713e-04	2.713e-04
Slop2	—	—	7.680e-08	7.680e-08
Residual standard error	7.067	4.806	4.083	4.083
Multiple R-squared	.7064	.7076	.92	.92
Correlation Coefficient	.8404692	0.7017397	.8404692	.8404692

- Here in above Table, We have implemented Problem Statement 1(Calorie $\sim$ Fat) using Least Square , RanSac , Cubic Fit and Gradient Descent.

- As examining Multiple R-squared in the above table ,Cubic Fit and Gradient Descent are suitable for the given Problem Statement.

References

[1] Nug,Andrew *Implementation of Gradient Descent Equation*, available at - http://class.coursera.org/ml-003/lecture/24

[2] statisticsfun *Implementation of Least Square Estimation*, available at https://www.youtube.com/watch?v=JvS2triCgOY

[3] https://www.cs.cmu.edu/ ggordon/10725-F12/slides/05-gd-revisited.pdf