Longitudinal Analysis Of Farm Animal Development

Shrimani Tundurwar
201774709

Supervised by Prof. Charles Taylor

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

August 2024

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity for Individual Pieces of Work

I declare that I am aware that as a member of the University community at the University of Leeds I have committed to working with Academic Integrity and that this means that my work must be a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine.

I declare that the attached submission is my own work.

Where the work of others has contributed to my work, I have given full acknowledgement using the appropriate referencing conventions for my programme of study.

I confirm that the attached submission has not been submitted for marks or credits in a different module or for a different qualification or completed prior to entry to the University.

I have read and understood the University’s rules on Academic Misconduct. I know that if I commit an academic misconduct offence there can be serious disciplinary consequences.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties to verify that this is my own work, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and I wish to have taken into account.

Student Signature:

Student Number: 201774709

Student Name: Shrimani Tundurwar

Date: 01/09/2024

Please note:

When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:

"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.

I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity."'

UNIVERSITY OF LEEDS

School of Mathematics

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Academic integrity statement

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes. I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name : Shrimani Tundurwar

Student ID : 201774709

Abstract

Over several years, a local farm in Leeds has meticulously collected data on newborn lambs, focusing on their growth patterns post-weaning. This data encompasses measurements of weight recorded monthly, alongside contextual information such as the lamb's sex, sire (father), dam (mother), breed, etc. The purpose of this research is to investigate the variation of growth curves based on several parameters, such as differences in growth based on litter size, the impact of sex on growth trajectories, and the influence of the breed of lamb on growth patterns. Understanding these variations is crucial for optimizing current farm practices and increasing profitability. By identifying the key factors that influence lamb growth, farmers can make informed decisions about breeding, feeding, and management practices to enhance the overall health and productivity of their flock. The approach begins with an exploratory data analysis to identify and rectify any anomalous values. Detailed longitudinal data analysis and the fitting of non-parametric and parametric models are then employed. By leveraging various statistical techniques, this study aims to provide a comprehensive understanding of the factors influencing lamb growth, ultimately offering valuable insights for improving farm practices and maximizing profitability.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The agricultural industry, particularly livestock farming, plays a crucial role in global food security and rural economies. Within this sector, the growth and management of sheep populations, particularly lambs, are of significant economic importance. Lambs are not only a valuable source of meat but also an important component of breeding programs aimed at improving the genetic quality of future generations. As a result, understanding and predicting lamb growth trends is crucial for improving farm management practices, improving profitability, and ensuring sustainable livestock production.

This dissertation focuses on the growth patterns of lambs, utilizing a comprehensive dataset collected from a local farm in Leeds over a seven-year period, from 2016 to 2022. The dataset includes detailed information on lamb growth measures, including weight, along with related variables like breed, gender, and litter size. Over these seven years, the farm has systematically recorded the development of lambs from birth to maturity, providing a rich source of longitudinal data for analysis.

1.1 Longitudinal Data Analysis

A key aspect of this dissertation is the longitudinal nature of the data, which allows for the study of growth patterns over time within the same cohort of lambs. Longitudinal data analysis is particularly useful since it makes it possible to identify temporal trends and evaluate growth trajectories [1]. It also provides a more in-depth understanding of how external factors, such as seasonal fluctuations or management practices, may impact growth over time. This method expands on longitudinal research by capturing the dynamic nature of growth, allowing for a more in-depth and accurate understanding of the factors influencing lamb development.

1.2 Information About Lambs

Under normal circumstances, the lifespan of lambs, or young sheep, can reach up to 10 to 12 years. However, in commercial farming, their life span is typically much shorter, depending on the purpose for which they are raised. The majority of lambs are raised primarily for meat production and are slaughtered within the first year of life, once they reach a certain target weight. On the farm from which the data for this study was collected, the targeted weight for butchering is 50 kg, which aligns with industry standards for lambs intended for meat.

The data gathered indicates that lambs on this farm are predominantly born in the month of March, which is a common practice in temperate regions. Scheduling the birth of lambs for March ensures that they arrive just as spring begins, when pasture conditions start to improve, offering abundant nutrition for the ewes and their young. After birth, lambs typically stay with their mothers and are weaned after a period of approximately 12 to 16 weeks [2]. Weaning is a critical stage in lamb development, as it marks the transition from a milk-based diet to a solid diet consisting mainly of pasture or other feed. The timing and management of weaning can have a significant impact on the growth rate and overall health of the lambs.

1.3 Research Objectives

The primary objective of this dissertation is to analyze the factors that influence lamb growth and to develop predictive models that can reliably estimate lamb weights based on the available data. By exploring how variables such as gender, breed, and litter size impact growth, this research aims to identify key trends and patterns that can inform better management decisions. Additionally, the study aims to examine the usefulness of various modeling approaches, both parametric and non-parametric, in predicting growth curves.

Specifically, the dissertation will address the following research questions:

1. How do gender, breed, and litter size affect the growth trajectories of lambs?

2. What are the effective statistical and machine learning models for predicting lamb weight at various stages of growth?

3. How can these models be used to improve farm management practices, particularly in terms of optimizing breeding strategies and maximizing meat production?

1.4 Significance of the Study

The findings from this research have the potential to contribute significantly to the field of livestock management. By providing a detailed analysis of growth patterns over a substantial period, this study can assist in finding optimal breeding procedures and management strategies that enhance both the quality and quantity of lamb production. The longitudinal study, in particular, offers a powerful framework for understanding how growth patterns develop over time, allowing for more precise predictions and targeted interventions.

This comprehensive approach ensures that the conclusions drawn are grounded in real-world observations, making them more applicable to practical farming situations.

Chapter 2

Data Description

2.1 Explaining the data

The dataset includes detailed records of the lamb's growth and development, including birth details, parental information, weaning data, and weights taken at various stages. Data on the sale or slaughter of lambs is also included, providing insights into their final live weight and the effects of various management approaches. By explaining each aspect of the data, this section aims to establish a foundation for the subsequent analysis and interpretation, ensuring that the context and relevance of the collected information are well understood.

2.1.1 Data Source

The dataset used in this study consists of detailed records of newborn lambs collected from a nearby farm in the city of Leeds (Cragg House Farm) over seven years. Monitoring and analyzing the growth patterns of lambs after they were weaned was the primary purpose of this data collection, with a focus on different variables that might have an impact on the development of the animals. The dataset, which covers the years 2016 through 2022, includes a variety of measurements and attributes associated with each lamb.

2.1.2 About The Data

The data were collected longitudinally, meaning that multiple observations were recorded for each lamb over time, allowing for a detailed analysis of growth patterns and other related factors.

Each lamb's weight was precisely recorded monthly at a random date rather than following a specific time interval, starting from weaning and continuing until the lamb was either sold or butchered. This method ensured a comprehensive growth record, yielding a robust dataset for longitudinal analysis.

Table 2.1 shows the summary of the data collected for the year 2020.

No.	Sex	D.O.B	Dam	Sire	Weaning Weight	Slaughter/Sale Date	DW	LW	Sold as Store or Breeding	23-07-2020	07-09-2020	07-10-2020	04-11-2020	03-12-2020	06-01-2021	03-02-2021	17-03-2021	14-04-2021	Type
821	M	17-Mar	443/16	1621	39	29-10-2020	22.5	48	Prime	39	45	49							Pedigree
822	F	17-Mar	443/16	1621	42	22-06-2021			Sulg ewe	42	44	50	55	54	55	54	55	62	Pedigree
823	M	18-Mar	690/18	2474	37	04-12-2020	28.6	57	Prime	37	47	52	52	57					Pedigree
824	F	18-Mar	690/18	2474	33					33	39	47	49	51	53	53	55	62	Pedigree
825	F	19-Mar	627/18	2474	33	22-06-2021			Sulg ewe	33	41	48	51	53	57	58	61	67	Pedigree
826	F	19-Mar	627/18	2474	33	17-09-2022			Breeding ewe	33	42	53	53	54	56	58	61	67	Pedigree
827	F	19-Mar	184/13	1621	42	22-06-2021			Sulg ewe	42	47	48	56	57	60	61	61	69	Pedigree
828	F	19-Mar	651/18	2474	37					37	43	45	52	52	55	55	59	65	Pedigree
829	F	19-Mar	651/18	2474	34					34	40	52	50	51	53	54	55	61	Pedigree
830	M	20-Mar	618/18	2474	30	22-01-2021	23.5	54	Prime	30	37	38	44	51	51				Pedigree
831	M	20-Mar	618/18	2474	31	14-01-2021	23.3	53	Prime	31	41	44	46	50	53				Pedigree
832	F	19-Mar	509/16	1621		13-04-2020			Pet Lamb										Pedigree
833	F	19-Mar	509/16	1621	41					41	40	45	50	51	53	54	55	61	Pedigree
834	M	19-Mar	509/16	1621	30	28-02-2021	22.1	50	Prime	30	37	40	45	50	49				Pedigree
835	F	20-Mar	654/18	2474	32	21-09-2021				32	40	45	49	49	53	54	55	64	Pedigree
836	M	20-Mar	654/18	2474	38	23-10-2020	24.7	53	Prime	38	46	50							Pedigree
837	M	21-Mar	479/16	1621	39	27-11-2020	28	60	Prime	39	48	52	56						Pedigree
838	F	21-Mar	685/18	2474	40	06-09-2021				40	45	50	54	55	58	58	61	68	Pedigree
839	F	21-Mar	510/16	1621	43	22-06-2021			Sulg ewe	43	45	52	58	59	60	61	60	67	Pedigree
840	M	21-Mar	553/17	1621	35	14-01-2021	26.2	53	Prime	35	42	45	48	53	53				Pedigree
841	M	21-Mar	553/17	1621	32	05-01-2021	25.6	50	Prime	32	37	42	46	52					Pedigree
842	F	21-Mar	553/17	1621	33	17-09-2022			Breeding ewe	33	39	44	50	49	53	54	57	65	Pedigree
843	M	21-Mar	1140/13	590	28	12-02-2021	22.1		Prime	28	39	42	44	49	49	52			Pedigree
844	M	21-Mar	1140/13	590	39	04-12-2020	27.6	57	Prime	39	46	50	53	57					Pedigree
845	M	22-Mar	444/16	1621	41	06-12-2021				41	49	51	55	59	58				Pedigree
846	F	22-Mar	649/18	2474	34					34	41	47	52	52	56	58	60	65	Pedigree
847	F	22-Mar	649/18	2474	32	22-06-2021			Sulg ewe	32	38	42	46	49	53	54	56	64	Pedigree
848	M	23-Mar	381/15	1621	45	24-09-2020	24.7	53	Prime	45	50								Pedigree
849	F	23-Mar	688/18	2474	33	22-06-2021			Sulg ewe	33	47	45	50	50	53	53	53	61	Pedigree
850	F	23-Mar	688/18	2474	36	21-09-2021				36	43	48	54	55	58	59	61	67	Pedigree
851	M	23-Mar	495/16	1621	38	05-02-2021	24.6	57	Prime	38	36	42	46	50	52	57			Pedigree
852	M	23-Mar	495/16	1621	34	14-01-2021	24.9	53	Prime	34	41	45	48	52	54				Pedigree
853	M	24-Mar	536/17	1621	38	26-02-2021	25	54	Prime	38	47	51	53	58	58				Pedigree
854	F	24-Mar	671/18	2474	33	17-09-2022			Breeding ewe	33	40	46	53	53	56	57	58	65	Pedigree
855	F	24-Mar	671/18	2474	30					30	38	43	47	48	51	51	52	61	Pedigree
856	M	25-Mar	362/15	1621		13-04-2020			Pet Lamb										Pedigree
857	M	25-Mar	362/15	1621		13-04-2020			Pet Lamb										Pedigree
858	M	25-Mar	362/15	1621	30	28-02-2021	23.9	50	Prime	30	39	42	45	47	47				Pedigree
859	M	25-Mar	459/16	1621	41	23-10-2020	24.7	53	Prime	41	50	51							Pedigree
860	M	25-Mar	459/16	1621	40	05-01-2020	24.3	50	Prime	40	47		50	52					Pedigree
861	F	25-Mar	617/18	2474	32	22-06-2021			Sulg ewe	32	38	42	46	48	47	48	49	58	Pedigree
862	F	25-Mar	617/18	2474	32					32	40		50	52	55	57	59	65	Pedigree
863	M	25-Mar	446/16	1621	36	26-03-2021			Prime	36	40	45	47	50	47	51			Pedigree
864	M	25-Mar	446/16	1621	28	22-01-2021	22.7	51	Prime	28	37	40	43	48	47				Pedigree
865	F	25-Mar	448/16	1621	41	22-06-2021			Sulg ewe	41	43	51	57	58	59	60	59	67	Pedigree
866	F	25-Mar	615/18	2474	31					31	39	45	49	51	56	56	57	64	Pedigree
867	M	25-Mar	615/18	2474	40	19-11-2020	26.6	57	Prime	40	46	50	54						Pedigree
868	M	26-Mar	542/17	1621	42	05-11-2020	28.4	60	Prime	42	53	56	60						Pedigree
869	M	26-Mar	542/17	1621	41	19-11-2020	27.8	58	Prime	41	46		55						Pedigree
870	M	27-Mar	455/16	1621	40	29-10-2020	22.9	49	Prime	40	47	51							Pedigree
871	F	27-Mar	455/16	1621	30	31-10-2022				30	38	43	50	47	48	47	50	58	Pedigree
872	M	27-Mar	677/18	2474	39	11-12-2020	25.2		Prime	39	44	47	50	56					Pedigree
873	F	27-Mar	679/18	2474	33					33	40	46	51	53	57	58	60	67	Pedigree
874	F	27-Mar	679/18	2474	31	22-06-2021			Sulg ewe	31	37	41	45	49	50	52	54	61	Pedigree
875	M	28-Mar	680/18	2474	30	12-02-2021	23.1		Prime	30	37	39	42	47	48	53			Pedigree
876	F	28-Mar	602/12	590	34	22-06-2021			Sulg ewe	34	39	45	51	52	54	53	53	61	Pedigree
877	F	29-Mar	323/15	1621	42	22-06-2021			Sulg ewe	42	48	53	58	58	60	60	62	69	Pedigree
878	F	29-Mar	323/15	1621	30					30	36	42	45	47	51	52	52	60	Pedigree
879	F	29-Mar	389/15	1621	34					34	39	45	50	48	51	51	54	62	Pedigree
880	M	29-Mar	389/15	1621	38	27-11-2020	26.4	56	Prime	38	45	49	51						Pedigree
881	M	02-Apr	1060/13	590	38	11-12-2020	26.8		Prime	38	46	51	51	56					Pedigree
882	M	02-Apr	1060/13	590	25	26-02-2021	21.1	45	Prime	25	32	39	43	45	47	48			Pedigree
883	M	02-Apr	526/16	1621	29	05-02-2021	23.9	55	Prime	29	37	43	45	50	49	55			Pedigree
884	M	02-Apr	526/16	1621	26	15-04-2021	28.4	59	Prime	26	32	36	40	46	47	52			Pedigree
885	M	02-Apr	526/16	1621	28	28-02-2021	23.9	50	Prime	28	35	49	41	47	49				Pedigree
886	M	05-Apr	328/15	1621	43	24-09-2020	24.5	52	Prime	43	51								Pedigree
800	M	19-Mar	613/18	Humphrey		14-05-2020			Sold with ewe										Commercial
430	F	19-Mar	613/18			14-05-2020			Sold with ewe										Commercial
428	M	20-Mar	303/15			13-04-2020			Pet Lamb										Commercial
187	M	21-Mar	550/17		45	24-07-2020	21.3	45	Prime	45									Commercial
429	M	21-Mar	550/17		41	24-07-2020	19.5	41	Prime	41									Commercial
188	F	22-Mar	580/17		38	07-08-2020	20.7	43	Prime	38									Commercial
189	F	22-Mar	580/17		36	21-08-2020	21.1	44	Prime	36									Commercial
190	F	23-Mar	616/18		21	05-03-2021	20.1	43	Prime	21	29	35	40	41	44				Commercial
595	M	23-Mar	616/18		31	09-10-2020	22.3	47	Prime	31	40	47							Commercial
596	F	23-Mar	616/18		31	08-09-2020	19.1	41	Prime	31	41								Commercial
597	F	26-Mar	641/18			16-05-2020			Sold with ewe										Commercial
598	F	26-Mar	641/18			16-05-2020			Sold with ewe										Commercial
599	M	26-Mar	520/16		31	16-10-2020	21.5	43	Prime	31	41	43							Commercial
600	M	26-Mar	520/16		35	21-08-2020	19.3	42	Prime	35									Commercial
801	F	26-Mar	520/16		27	05-11-2020	22.3	43	Prime	27	35	39	43						Commercial
802	F	26-Mar	1147/13		26	16-10-2020	19.7	42	Prime	26	33	40							Commercial
803	M	26-Mar	1147/13		39	27-08-2020	18.7	40	Prime	39									Commercial
804	F	27-Mar	1061/13		35	02-10-2020	22.5	50	Prime	35	41								Commercial
805	M	27-Mar	1061/13		27	09-10-2020	21.7	45	Prime	27	39	45							Commercial
806	M	27-Mar	1061/13		36	08-09-2020	19.7	44	Prime	36	44								Commercial
807	M	27-Mar	621/18			19-05-2020			Sold with ewe										Commercial
808	M	27-Mar	621/18			19-05-2020			Sold with ewe										Commercial
809	F	29-Mar	647/15			16-05-2020			Sold with ewe										Commercial
810	F	29-Mar	543/17			14-04-2020			Pet Lamb										Commercial
811	F	29-Mar	543/17		29	02-10-2020	21.3	45	Prime	29	40								Commercial
812	F	29-Mar	543/17		34	27-08-2020	18.3	39	Prime	34									Commercial
813	M	06-Apr	1153/13		39	07-08-2020	20.9	45	Prime	39									Commercial

Table 2.1: The table shows the yearly data gathered for 2020, including both pedigree and commercial lamb breed.

The comprehensive and systematic approach to data collection, along with the inclusion of detailed contextual information, provides a robust dataset for analyzing lamb growth patterns. The longitudinal nature of the data provides for a detailed understanding of how various parameters such as sex, weaning weight, and breed influence lamb development over time.

Contextual Information

At the start of the data collection process, extensive initial information was recorded for each lamb. This included:

- **D.O.B:** The birth date of each lamb was documented to calculate the age at each measurement point accurately.
- **Dam & Sire:** The genetic background of each lamb was captured by recording the identification numbers and relevant details of both the sire (father) and dam (mother). This information is crucial for understanding the genetic influences on growth patterns.

Subsequent Contextual Data

In addition to monthly weight measurements, various contextual data points were recorded throughout the lamb's life, including:

- **Slaughter/Sale Date:** The date on which each lamb was either sold or butchered was recorded to determine the period of data collection for each individual.
- **Sold as Store or Breeding:** If a lamb was sold, the purpose of the sale was specified, whether the lamb was sold as a pet, for breeding, or for meat production. This information helps to understand the end use of the lambs and any potential impact on their growth patterns.
- **Type:** The type of lamb representing its breed was recorded.

Additional Information

- **Weaning Weight** : The weaning weight is the recorded weight of a lamb on the exact date when the lamb transitions from relying on its mother's milk to an independent diet, typically composed of solid feed or pasture.
- **DW:** The dead weight refers to the weight of a lamb's carcass after slaughter but before the removal of non-carcass components such as the head, hide, hooves, and internal organs.
- **LW:**The live weight at the time of sale or slaughter refers to the final recorded weight of the lamb just before it is sold or sent for processing. This weight, measured in kilograms, provides a critical snapshot of the lamb's overall growth and condition at the endpoint of its life.

Variable Name	Type	Description
D.O.B	Date	The precise birth date of each lamb, used to calculate age at each measurement point.
Dam	Categorical	Identification number of dam (mother) to capture genetic background.
Sire	Categorical	Identification number of sire (father) to capture genetic background.
Slaughter/Sale Date	Date	The date when each lamb was either sold or butchered, marking the end of data collection for that individual.
Sold as Store or Breeding	Categorical	Specifies the purpose of the sale—whether the lamb was sold as a pet, for breeding, or for meat production.
type	Categorical	The breed of the lamb, representing its genetic type.
Weaning Weight (Kg)	Numeric	The weight of the lamb at the time of weaning, a critical developmental milestone.
DW (Kg)	Numeric	The weight of the lamb's carcass after slaughter but before removal of non-carcass components.
LW (Kg)	Numeric	The final recorded weight of the lamb just before it is sold or sent for processing.

Table 2.2: Summary of Variables in the Lamb Growth Dataset.

2.2 Data Preparation

Data preparation is a crucial step in longitudinal data analysis, particularly when dealing with complex datasets such as lamb growth data collected over several years. The process involves cleaning, transforming, and structuring the data to make it suitable for accurate and relevant analysis. Given that the data spans seven years, there are inconsistencies such as varying date formats, missing weight records, and potential errors. To address these issues, the data preparation process includes the following key steps:

- **Data Formatting**
- **Data Transformation**
- **Missing Data**
- **Redundant Data**

This section outlines the iterative process of finding and addressing incomplete, inaccurate, and irrelevant data as part of the CRISP data mining framework (see Figure 2.1 [3]).

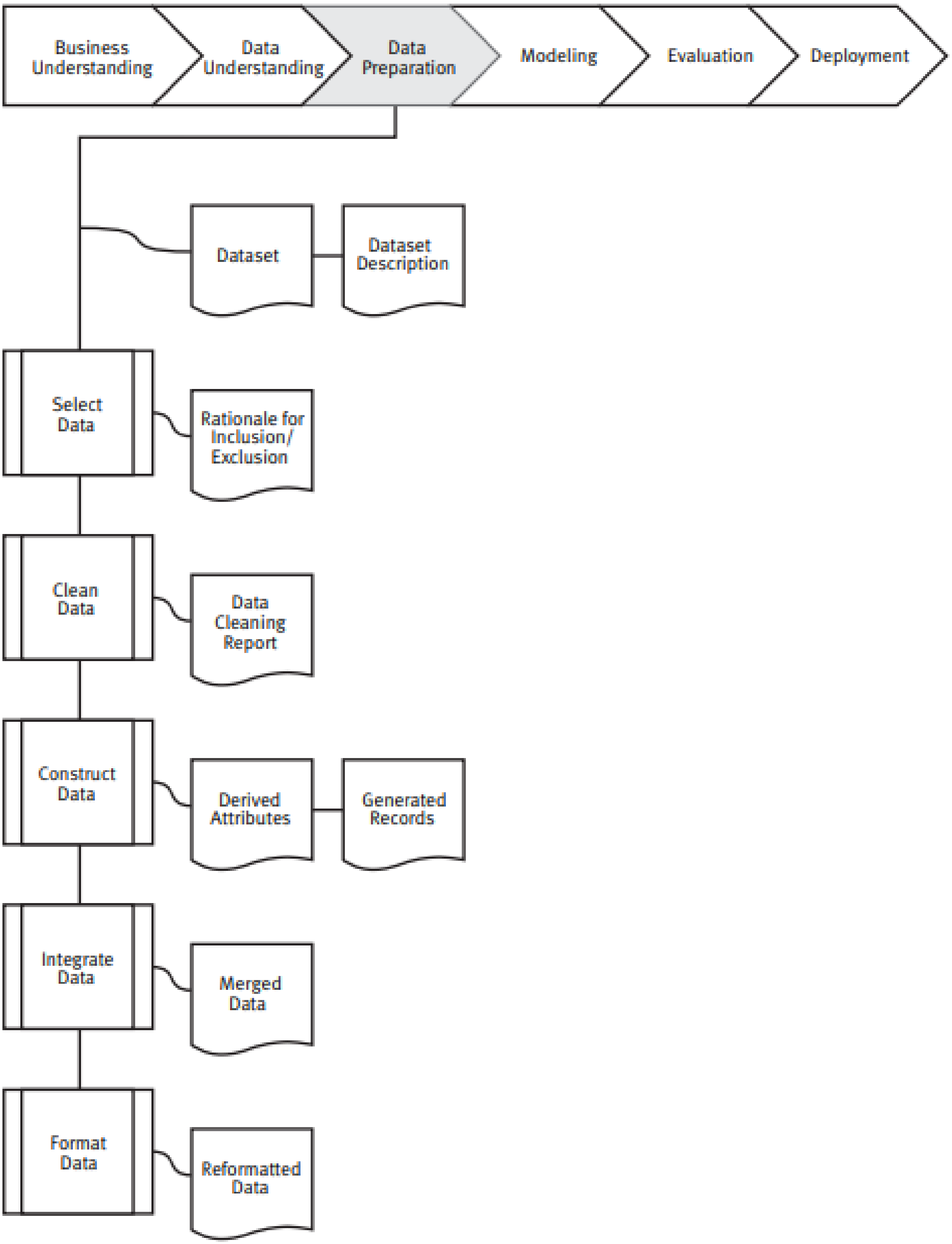
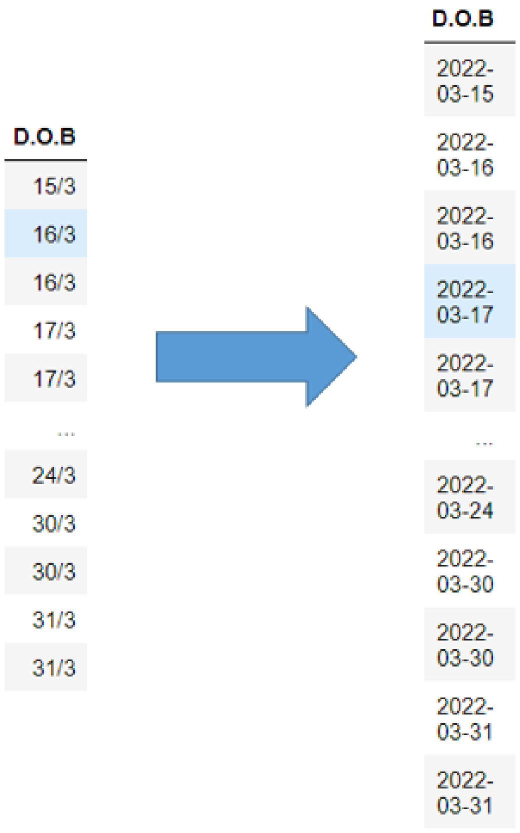


Figure 2.1: The data preparation stages of data analysis.

2.2.1 Data Formatting

In the dataset, proper data formatting is crucial to prevent issues when merging data from different years. One important aspect is the D.O.B. (Date of Birth) column, which contains only the day and month of birth. To avoid redundancy and ensure accuracy, it is essential to append the correct year to these values. This step is critical for maintaining consistency when integrating data across different time periods.

Additionally, all date-related columns, including the Slaughter/Sale Date and various weight measurement dates, must be converted to appropriate date formats. Ensuring that these columns are stored as appropriate date types is necessary for accurate data processing, such as performing date comparisons and time-series analysis. These changes are critical for the analysis and data processing requirements discussed in subsequent sections.



D.O.B
15/3
16/3
16/3
17/3
17/3
...
24/3
30/3
30/3
31/3
31/3

D.O.B
2022-03-15
2022-03-16
2022-03-16
2022-03-17
2022-03-17
...
2022-03-24
2022-03-30
2022-03-30
2022-03-31
2022-03-31

Figure 2.2: Formatted data where year was introduced in the D.O.B column.

2.2.2 Data Transformation

Data transformation is also an important stage in data analysis, as it converts raw data into a more suitable format for study. This process involves modifying, aggregating, or deriving new variables from existing data to better align with the objectives of the analysis. By transforming the data, meaningful patterns, trends, and relationships can be uncovered that may not be immediately apparent in the raw data.

The weights of each lamb in the dataset are recorded on specific dates. To facilitate longitudinal data analysis, it is necessary to transform this information into the age of the lamb at the time of each weight measurement. This transformation allows it to analyze the growth patterns of each lamb over time rather than just looking at isolated weight measurements. For example, each lamb will have multiple rows of data representing its weight at specific ages, enabling a more detailed analysis of its growth curve and developmental trajectory.

To calculate the age of a lamb, the difference between the lamb's date of birth and the date of each recorded weight measurement was used. Mathematically, this is expressed as:

$$\textbf{Age of Lamb (Days)} = D_w - D_b$$

Where:

- **Date of Weight Measurement** (D_w): The date on which the lamb's weight was recorded.
- **Date of Birth (D.O.B.)** (D_b): The date when the lamb was born.

The difference between these two dates provides the lamb's age at the time of each weight measurement, typically expressed in days. This age value serves as a crucial variable for longitudinal analysis, which allows it to trace the growth of each lamb over time.

The date columns in the dataset that were converted to lamb age were transposed, resulting in multiple rows for each lamb. Each row indicates a specific weight measurement, including the lamb's age and weight on the given date. This transition provides a precise picture of the lamb's development over time.

Lamb No.	Age (Days)	Weight
186	103	38.0
186	130	45.0
186	157	45.0
186	193	NaN
186	223	NaN
...	...	...
5820	328	NaN
5820	358	NaN
5820	363	NaN
5820	383	NaN
5820	386	NaN

Table 2.3: Summary of transformed data from the original data source.

The initial dataset of 578 rows was expanded to 6,178 rows due to the transformation, with each weight measurement entry converted into a separate row to include detailed age and weight information.

2.2.3 Missing Data

In longitudinal studies, where data is collected from the same subjects over time, missing data is a common issue that can complicate the analysis. In the context of lamb data, which tracks the weight of individual lambs over various days, missing weight measurements (indicated by ”NaN”) can occur due to various reasons such as measurement errors, logistical challenges, or health issues that prevent the collection of data at certain time points.

These missing data points can introduce bias and reduce the statistical power of the analysis if not properly addressed. It is crucial to consider appropriate methods for handling missing data, such as imputation techniques, which can provide robust estimates even when some data points are missing [4]. Addressing missing data ensures that the conclusions drawn about growth patterns and other longitudinal trends in the lambs are accurate and reliable.

In this dataset (refer to tables 2.3 and 2.1), the issue of missing data extends beyond just the age or weight measurements of the lambs. Critical information such as sire and dam details, live and dead weights, weaning weights, and even records of when the lambs were butchered or sold are also missing in some cases. This incomplete data can pose significant challenges for longitudinal analysis. Missing sire and dam information makes it difficult to evaluate genetic implications on growth patterns, while missing weight records at different stages, such as weaning or live weights, can obscure key stages of development. Likewise, it is challenging to determine the precise timing of these events in relation to the growth of the lamb because there are no timelines for sales or slaughtering.

Handling this incomplete data is essential in ensuring the accuracy of any conclusions drawn from the analysis. One approach could involve the use of multiple imputation techniques, which can fill in missing values based on the patterns observed in the available data. It is also important to consider the possibility of data being missing not at random (MNAR), where the likelihood of data being missing is related to the missing values [5]. For instance, in this dataset, the missing weights of lambs on certain observation dates may not be random but rather indicative of specific events, such as the lamb being sold or butchered. This type of missing data can have a significant impact on the analysis because the absence of values is tied to a meaningful underlying reason that must be accounted for in order to avoid biased conclusions. Understanding and addressing these various types of missing data will be crucial in drawing reliable conclusions about the growth, development, and management outcomes of the lambs in the study. In the subsequent sections, missing data will be addressed during the modeling phase.

2.2.4 Redundant Data

Redundant data in longitudinal analysis can be difficult to deal with, especially in studies that involve repeated measurements over time. Redundancy arises when multiple observations provide the same or similar information, resulting in a bias of particular data points. This can happen when measurements are too frequent or when there is little variation between successive observations. In such cases, the redundancy can inflate the dataset, making it more complex without adding valuable insights. This type of data might cover up natural patterns and relationships, thereby skewing results and lowering the effectiveness of statistical models. Addressing these often involves careful consideration of the frequency and relevance of data collection to ensure that each observation adds value to the study. Reducing redundant data helps simplify the dataset, enhance interpretability, and improve the accuracy of longitudinal studies.

To ensure the integrity of the dataset, all instances of redundant data were carefully checked and removed, eliminating any duplicate records that could skew the analysis. This meticulous process ensured that each data point contributed uniquely to the study, enhancing the accuracy and reliability of the longitudinal analysis.

2.3 Data Exploration

The dataset serves as the foundation for analyzing the growth patterns of lambs, a vast collection of variables that provide a comprehensive view of each lamb's development from birth to either sale or slaughter. This section delves into the key components of the dataset, focusing on their importance in understanding the lambs' growth trajectories and overall health.

The significance of this dataset lies in its ability to capture a wide range of factors that influence lamb growth, including genetic background and management practices. By examining the birth records, parental information, weaning data, and subsequent weight measurements, it is possible to identify patterns and correlations that may help to optimize lamb growth rates and improve flock management overall.

This dataset, in particular, allows for the tracking of live weights at various stages, providing insights into how different factors, such as breed and parental characteristics, impact growth. Furthermore, the data on slaughter or sale outcomes offers a practical view on how these growth patterns translate into economic returns, whether through meat production or breeding stock sales.

2.3.1 Impact of different factors

Identifying the factors that have a major impact on the growth curve is crucial for understanding the reasons for growth patterns. In this section, the effects of variables like weaning weight, gender, and type on lamb growth have been evaluated.

Type of Lamb (Breed)

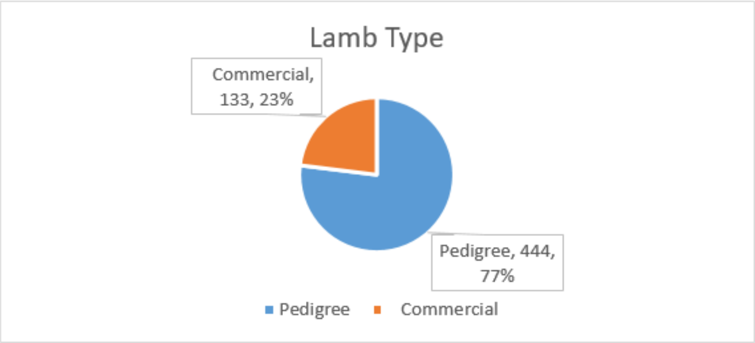


Figure 2.3: The pie chart shows the lambs division between pedigree and commercial breed.

The dataset comprises 578 lambs, categorized into two distinct types: **'pedigree'** and **'commercial'**. Pedigree lambs dominate the distribution, representing 444 of the total, which is approximately 77%, while commercial lambs account for 134, or about 23%. The primary focus on pedigree lambs shows a desire to preserve and improve specific genetic lineages, possibly for breeding or premium meat production.

Year	Pedigree	Commercial	Total
2016	99	0	99
2017	15	48	63
2018	80	0	80
2019	69	17	86
2020	66	27	93
2021	62	21	83
2022	53	21	74

Table 2.4: Distribution of Pedigree and Commercial Lambs Over the Period of Seven Years.

As depicted in Table 2.4, particularly in 2017, the majority of lambs were commercial (48 out of 63), which contrasts with other years where pedigree lambs were consistently in the majority. This anomaly in 2017 indicates a shift in breeding or sales focus for that year, which could be due to market demands, breeding plans, or other external reasons influencing the commercial to pedigree lamb ratio. In all other years, pedigree lambs are the most common, indicating a general trend of maintaining or increasing pedigree lines.

Furthermore, the data supports the hypothesis that commercial lambs have a significant impact on the growth curve. As seen in Table 2.5, a higher proportion of commercial lambs are butchered and sold as meat products, possibly due to their greater mass. This trend highlights the impact of breed type on lamb growth and market outcomes.

Years	**Pedigree**		**Commercial**	
	Butchered	**Not - Butchered**	**Butchered**	**Not - Butchered**
2016	32	67	0	0
2017	11	4	22	26
2018	34	46	0	0
2019	27	42	17	0
2020	33	33	18	9
2021	17	45	19	2
2022	16	37	21	0

Table 2.5: Annual distribution of pedigree and commercial lambs, categorized by their status as butchered or not butchered, from 2016 to 2022.

The data presented in Table 2.5 highlights distinct differences in the utilization of pedigree and commercial lambs over the years 2016 to 2022. The commercial lambs, as the table indicates, are mostly butchered, reflecting their primary role in meat production. This trend suggests that commercial lambs are generally raised for their larger mass, making them more suitable for butchering and sale as meat products.

Comparatively, the pedigree lambs have a different pattern of use. A relatively small number of pedigree lambs are butchered, while the majority are kept for breeding or sold as pet or spare lambs. This distinction highlights the several advantages of pedigree lambs, which are often valued not just for their meat but also for their genetic qualities and versatility. As a result, pedigree lambs are more frequently kept alive to capitalize on these advantages, whether in breeding programs or other roles.

Weaning Weight of Lambs

Weaning weight, which indicates a lamb's weight at the time of separating from its mother's milk, is an important production statistic. This weight, which is impacted by a variety of factors including genetics, mother care, food, and general management practices, represents the health and growth of the lamb during its early life. Weaning weight affects a lamb's future growth and market value while also providing an indicator of its early development potential.

The analysis of weaning weights from 2016 to 2022 reveals an average weight of **37.62 kg** across all lambs. During this period, the recorded weaning weights varied significantly, with the highest reaching **55 kg** and the lowest at **21 kg**. These figures highlight the range of growth outcomes observed during this crucial developmental stage.

The data shows that commercial breed lambs averaged **39.61 kg** at weaning, while the average weight of pedigree lambs was **37.78 kg**. This discrepancy suggests that compared to their pedigree counterparts, commercial lambs typically weigh heavier. According to these results, the **breed** of the lamb has a significant effect on the growth trajectory of lambs, with commercial breeds appearing to have an edge in the early stages of growth.

Years	Average Weaning Weight	Max Weaning Weight Recorded	Minimum Weaning Weight Recorded
2016	40.02	55	28
2017	44.73	55	26
2018	34.05	46	21
2019	35.99	46	24
2020	34.81	45	21
2021	36.63	47	26
2022	38.80	50	26

Table 2.6: Yearly Overview of Weaning Weights: Average, Maximum, and Minimum Values from 2016 to 2022.

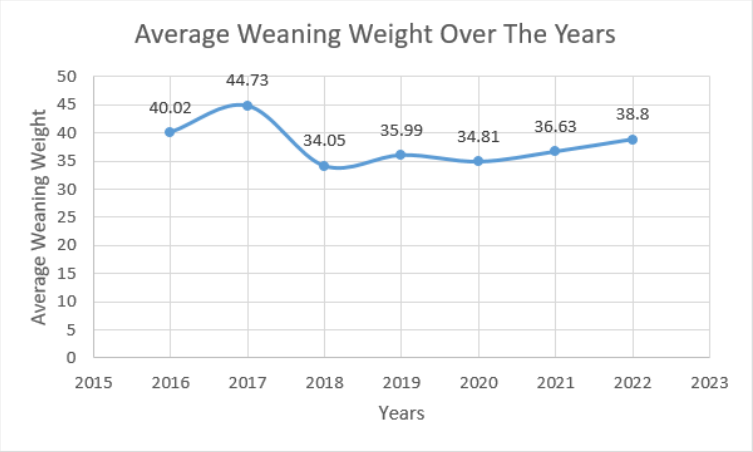


Figure 2.4: Trend of average weaning weight over the years of 2016-2022.

Figure 2.4 shows fluctuations in the average weaning weight from 2016 to 2022. Notably, there is an increase in average weaning weight in 2017, peaking at 44.73 kg, before experiencing a decline in the following years. The slight upward trend could be attributed to a higher proportion of commercial lambs introduced in 2017 (see table 2.4). Commercial lambs often have greater weaning weights than pedigree lambs, which may have contributed to the overall increase in average weaning weights seen in 2017, with the average weight increasing slightly each year thereafter.

Gender Based Analysis

Gender plays a significant role in the growth performance of lambs, with male and female lambs often exhibiting distinct growth patterns due to physiological and hormonal differences. Male lambs typically grow faster and achieve higher body weights compared to females. In contrast, female lambs may mature earlier but with slower growth rates. Analyzing these gender-based differences in lamb growth is essential for developing targeted management strategies, optimizing productivity, and ensuring more efficient livestock production. This section explores the impact of gender on lamb growth curves, highlighting the importance of considering gender in sheep farming practices.

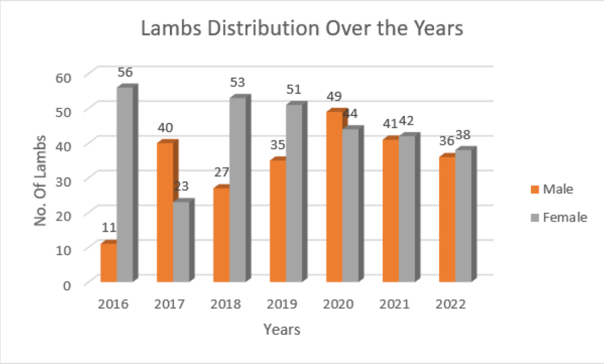


Figure 2.5: Lamb distribution based on gender over the years 2016 to 2022.

Figure 2.5 indicates significant gender differences in the number of male and female lambs across various years. In 2016, female lambs outnumbered males by about 5:1 (56 females to 11 males). In successive years, such as 2017, 2018, and 2019, the female count remained continuously higher, nearly double that of males, demonstrating a significant gender disparity. From 2019 to 2022, the number of male and female lambs became more balanced, with both genders showing similar counts.

This gender distribution is crucial in understanding the fluctuations in average weaning weights across the years, as male lambs generally have higher weaning weights than females (refer table 2.7). The years with a higher proportion of female lambs could experience lower overall average weaning weights due to the lower average weight of female lambs (see table 2.6). Conversely, years with a higher male lamb distribution might see increased average weaning weights. Therefore, analyzing the male percentage distribution over the years is essential to predict and understand the potential variations in average weaning weights.

Years	**Male**	**Female**	**Combined Average Weaning Weight**
2016	48.00	37.68	40.02
2017	46.28	42.04	44.73
2018	37.55	32.26	34.05
2019	38.03	34.59	35.99
2020	35.63	33.89	34.81
2021	39.19	34.63	36.63
2022	39.39	38.20	38.80

Table 2.7: Average Weaning Weights of Male and Female Lambs from 2016 to 2022.

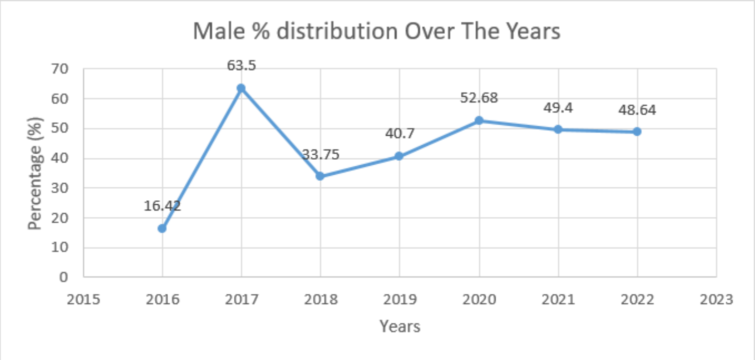


Figure 2.6: Figure shows the percentage distribution of male lambs over the period of 2016-2022.

Figure 2.6 clearly shows that the percentage of male lambs in 2017 was significantly higher compared to other years, which has a noticeable impact on the average weaning weight curve (see figure 2.3). This increased male dominance in 2017 contributed to the peak observed in the average weaning weight, as male lambs typically have higher weaning weights than females. Conversely, in 2016, the percentage of male lambs was the lowest, comprising only 16% of the total lambs. This minimal representation of males in 2016 likely contributed to the relatively lower average weaning weight observed for that year. These gender distribution variations are key factors in understanding the fluctuations in the average weaning weight over the years.

The data suggests that sex plays a significant role in shaping the growth curve of lambs, with male lambs consistently showing higher weaning weights compared to females. This correlation indicates that leveraging the knowledge of the higher weaning weights of male lambs could be crucial for optimizing growth predictions and management strategies in sheep farming, ultimately influencing the overall productivity and efficiency of the operation.

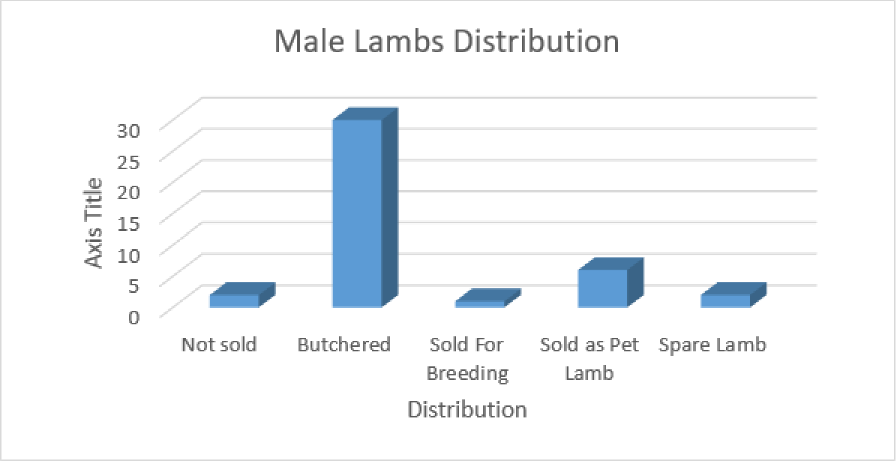


Figure 2.7: The figure shows the male lamb sold or butchered in the year 2016.

Figure 2.7 from 2016 illustrates the distribution of male lambs across various categories, including those butchered, sold as pets, kept as spare lambs, or used for breeding. It shows that a significant portion, around 90%, of lambs were butchered and sold as meat products. This pattern is seen over the whole dataset and is not limited to 2016. It suggests that producing meat has been the primary goal of raising male lambs. Year after year, a similar proportion, around 90%, of male lambs were directed towards butchering, with only a small fraction allocated to other uses such as pets, spare lambs, or breeding. This recurring trend highlights the primary role of male lambs in the meat industry.

2.3.2 Growth Curves

Growth curves are essential tools for understanding the development of lambs over time, particularly in terms of their weight and size. These curves graphically represent how lambs grow from birth through various stages of development, providing insights into their growth patterns. By analyzing growth curves, farmers can identify key phases of rapid growth and periods of slower development. This information is crucial for evaluating the effectiveness of feeding schedules, health interventions, and overall management practices. Deviations from expected growth patterns may indicate nutritional deficiencies or health issues that need addressing. Additionally, understanding growth curves helps in making informed decisions about the optimal timing for selling or butchering lambs, ultimately aiding in maximizing profitability and achieving desired market weights.

To generate the growth curves, Python was utilized with the **'matplotlib'** library, providing a clear visual representation of lamb growth data over time.

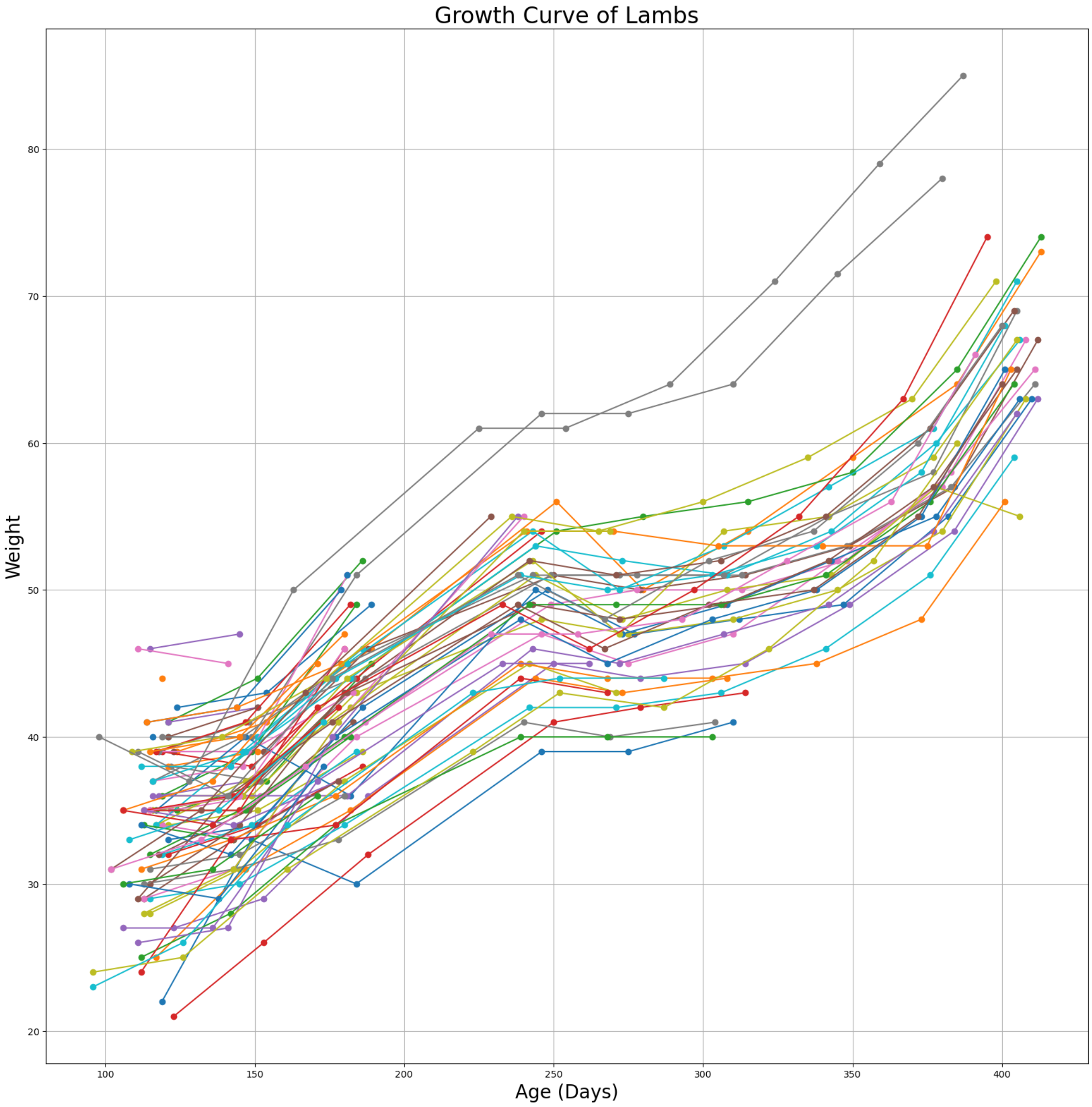


Figure 2.8: The figure shows the typical growth curves of lambs in the year 2018.

Figure 2.8 presents a detailed visualization of the growth curves of multiple lambs over time for the year 2018, with each curve representing an individual lamb's weight as it ages. The x-axis marks the age of the lambs in days, spanning from approximately 100 to 400 days, while the y-axis represents their weight, ranging from around 20 to 90 kg. The plot shows a collection of growth trajectories, with each curve drawn in a different color to represent the variations in patterns amongst individual lambs.

The data points, marked by small circles, indicate the recorded weights of the lambs at specific ages. These points are connected by continuous lines, providing a clear view of the growth trend for each lamb, even when there are missing data points. The continuous nature of these lines ensures that the overall growth pattern is maintained, making it easier to track the progression of weight over time despite any gaps in the data.

Overall, the figure depicts a general upward trend in the lambs' growth, with most curves showing an increase in weight as the lambs age with a potential fluctuation around the age, of 250 days. However, there is notable variability in the growth rates, with some lambs experiencing rapid weight gain while others show a more gradual increase.

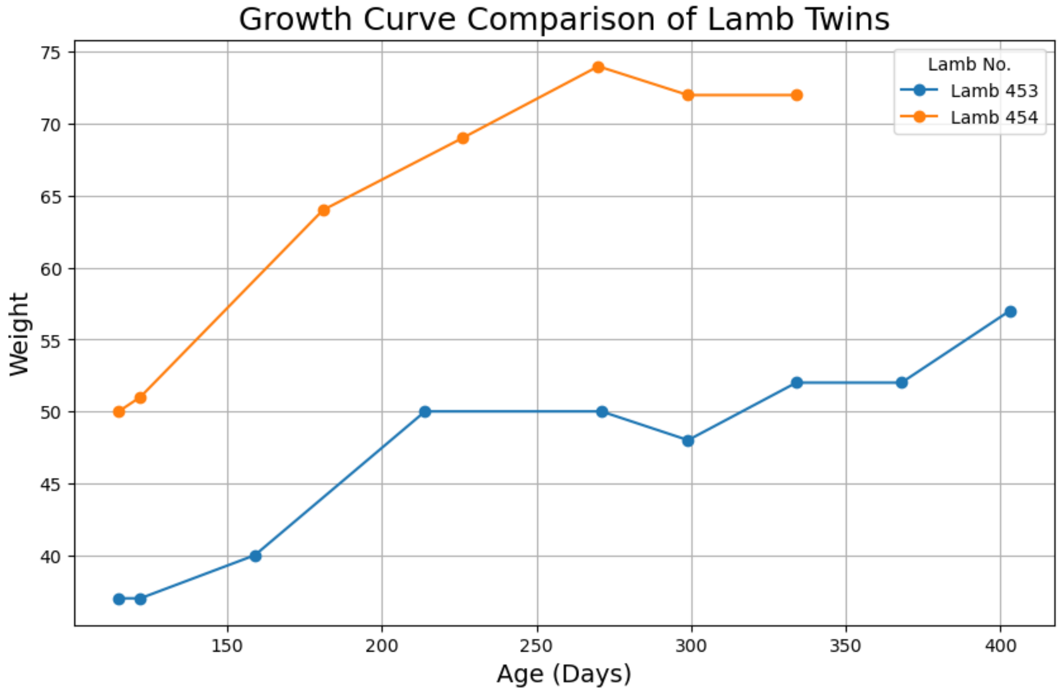


Figure 2.9: The figure shows the comparison between the growth curves of two lambs, male (454) and female (453) from the year 2018.

Figure 2.9 shows a comparison of growth curves between two lambs, Lamb 453 and Lamb 454, which are siblings born from the same mother. Lamb 454, a male lamb, shows a significantly steeper growth curve compared to Lamb 453, a female lamb. Over time, Lamb 454 consistently gains weight at a faster rate, reaching a weight of approximately 75 kg by the age of 300 days, after which its weight stabilizes.

However, in comparison, Lamb 453 exhibits a more gradual increase in weight, with periods of stagnation and even a slight decline around 300 days. By the age of 400 days, Lamb 453's weight remains below 55 kg, significantly lower than that of its male sibling.

Additionally, the graph for both of the lambs shows some missing data points, reflecting instances where weight measurements were not recorded for the lamb on that specific day.

This comparison highlights the differences in growth patterns between male and female lambs, suggesting that gender plays a role in the growth rate and overall size of the lambs. Despite being born from the same mother, the male lamb's growth is more robust, indicating a stronger response to growth stimuli.

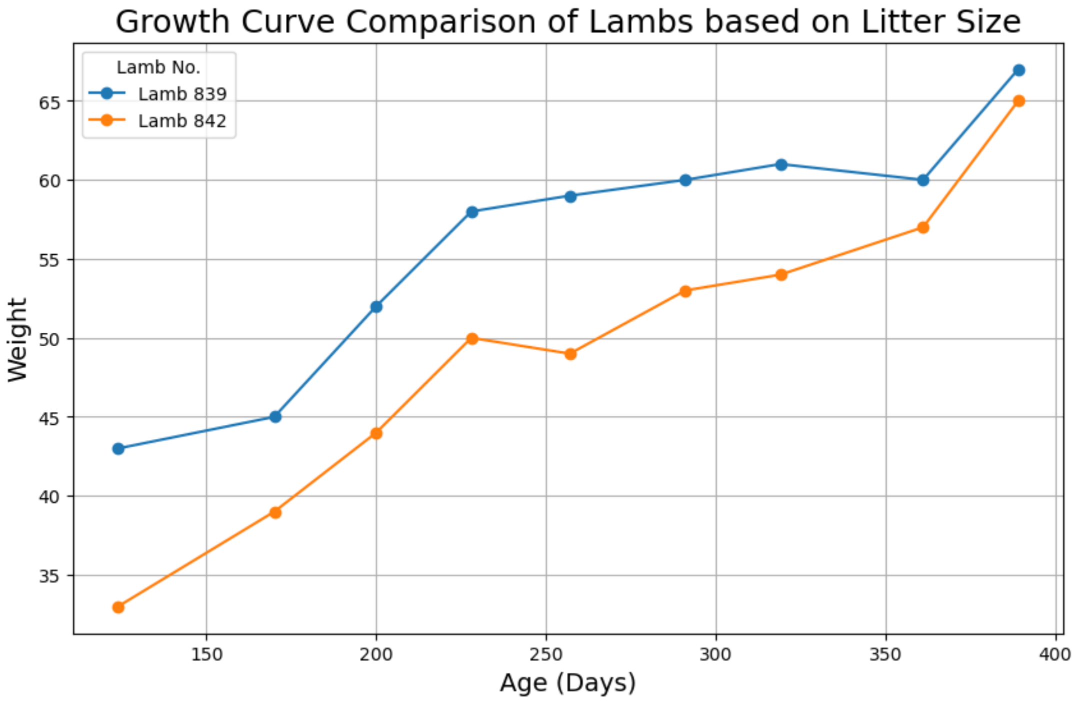


Figure 2.10: The figure shows the comparison between the growth curves of two lambs (828,829) both female.

Figure 2.10 shows that, despite being females, lambs 839 and 842 had very different growth curves. This variation highlights the impact of litter size on growth patterns, with lamb 842 being a twin and lamb 839 being a single birth. The difference in their growth curves demonstrates how lambs can have different developmental trajectories based on their birth order and litter size.

Following the principle of growth curves, this implies the necessity of developing a model to estimate growth trajectories based on the specific data for the lambs, which will be discussed in the following chapters.

Chapter 3

Methodology

This chapter introduces the statistical models and techniques employed to analyze the growth patterns of lambs. The analysis relies on a variety of methodologies appropriate for the data, implemented using tools such as Microsoft Excel and Python libraries.

To effectively capture the growth trajectories observed, both flexible non-parametric methods and more structured parametric techniques are used. These methods are selected based on their ability to handle the specific characteristics of longitudinal data, allowing for the precise estimation of growth patterns without relying on predefined assumptions.

The analysis focuses on identifying the most appropriate approach for predicting lamb growth and understanding the factors that influence it. By leveraging the capabilities of Excel and Python, the selected models provide a robust framework for exploring the dataset, leading to a deeper understanding of the growth dynamics and enhancing the accuracy of growth predictions.

3.1 Non-Parametric Models

Non-parametric models are essential tools in longitudinal data analysis, particularly when the underlying relationship between variables is unknown or complex. Unlike parametric models, which assume a specific form for the relationship between independent and dependent variables, non-parametric models do not impose any predefined structure on the data. This flexibility makes them highly effective for capturing discrete patterns, allowing variability, and adapting to the unique characteristics of each dataset.

3.1.1 The Importance of Non-Parametric Models in Longitudinal Data Analysis

Traditional parametric techniques, which collect data from the same subjects over time, may fail to capture the complex and individual-specific patterns of change. For example, in the context of a lamb dataset, the growth trajectories of individual lambs might differ significantly due to genetic, environmental, and management factors. Applying a parametric model might oversimplify these differences by imposing a specific functional form across all individuals.

Non-parametric models, by contrast, can estimate the growth curve for each individual without requiring the entire population to follow the same pattern. This is achieved through methods like Kernel Smoothing, Splines, and Gaussian Processes, which allow the data itself to guide the shape of the estimated function [6].

3.1.2 Mathematical Representation and Application

Consider a longitudinal dataset

$$\{(t_{ij}, y_{ij}) : i = 1, \ldots, N, \; j = 1, \ldots, T_i\},$$

where t_{ij} represents the time point of observation j for individual i, and y_{ij} is the corresponding response variable (e.g., weight of the lamb).

In a non-parametric framework, we do not assume that the relationship between y_{ij} and t_{ij} follows a specific parametric form like

$$y_{ij} = f(t_{ij}; \theta) + \epsilon_{ij},$$

where θ represents the parameters to be estimated and ϵ_{ij} denotes the error term. Instead, we model:

$$y_{ij} = f_i(t_{ij}) + \epsilon_{ij},$$

where $f_i(t)$ is an individual-specific smooth function, often estimated using techniques such as:

- **Kernel Smoothing**: The function $f_i(t)$ is estimated by weighting nearby points, where the weights are determined by a kernel function $K(\cdot)$. The estimate at time t is given by:

  $$\hat{f}_i(t) = \sum_{j=1}^{T_i} K\left(\frac{t - t_{ij}}{h}\right) y_{ij},$$

  where h is a bandwidth parameter controlling the smoothness [7].

- **Spline Regression**: The function $f_i(t)$ is represented as a piecewise polynomial, with the pieces smoothly connected at specific points called knots. The general form is:

  $$f_i(t) = \sum_{k=1}^{K} \beta_{ik} B_k(t),$$

  where $B_k(t)$ are basis functions (e.g., B-splines) and β_{ik} are the coefficients estimated from the data.

- **Gaussian Processes**: Here, $f_i(t)$ is modeled as a sample from a Gaussian process with a specified covariance function. This approach provides a probabilistic framework, allowing for the incorporation of uncertainty in predictions.

In the lamb dataset, where growth may vary significantly between individuals due to various factors, a non-parametric approach like the Nadaraya-Watson (NW) kernel regression estimator can be particularly effective. The Nadaraya-Watson estimator provides a smooth estimate of the growth trajectory by weighting nearby observations more heavily, without imposing a rigid functional form on the data. This makes it ideal for capturing the diverse and nonlinear growth patterns seen in individual lambs, accommodating the substantial inter-individual variability observed in the dataset.

3.1.3 Nadaraya-Watson Estimation Technique

The Nadaraya-Watson Estimation technique is a powerful non-parametric method utilized in statistical analysis to estimate the relationship between a dependent variable and an independent variable. This approach is particularly useful in cases where the relationship between the variables is complex or unknown, as it does not assume a specific functional form. Instead, it allows the data to determine the shape of the relationship, making it ideal for longitudinal studies such as analyzing the growth patterns of lambs over time.

Application in the Lamb Data

In lamb growth analysis, the Nadaraya-Watson estimator is used to estimate the weight of a lamb at a given age, even if the lamb's weight was not explicitly measured at that exact age. This is particularly valuable in longitudinal data analysis, where measurements may be irregular and the exact time points of interest may not coincide with the observed data points.

The technique works by assigning weights to each observed data point based on its proximity to the target point (e.g., the specific age at which the weight estimate is desired). These weights are determined using a kernel function, which ensures that data points closer to the target age have a greater influence on the estimated value than those further away.

Mathematical Formulation

The Nadaraya-Watson estimator for predicting the weight $\hat{y}(x_0)$ of a lamb at a target age x_0 is mathematically expressed as:

$$\hat{y}(x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right)},$$

where:

- x_i and y_i represent the observed ages and corresponding weights of the lambs, respectively.
- $K(\cdot)$ is the kernel function, which determines the weights for the data points [8]. A common choice for the kernel function is the Gaussian kernel, given by:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}.$$

- h is the bandwidth parameter that controls the smoothness of the estimated curve. A smaller bandwidth makes the estimator more sensitive to local fluctuations in the data, while a larger bandwidth results in a smoother curve by averaging over a broader range of data points.

The Nadaraya-Watson estimator is implemented using computational tools such as Python, which allow for efficient handling of large datasets and the computation of weighted averages. By applying this method to the lamb growth data, it is possible to generate a smooth growth curve that captures the overall trend in weight changes over time, as well as allowing individual variations in growth patterns. This smooth curve can then be used to predict the weight of a lamb at any age, filling in gaps between observed measurements and providing a continuous representation of growth.

The Nadaraya-Watson Estimator was implemented to analyze lamb growth patterns by utilizing several key functions which were manually created in the code:

- The **'gaussian_kernel'** function calculates Gaussian kernel weights, which determine how much influence each observed data point (age) has on the estimated weight at a target age. This function uses the in-built 'scipy.stats.norm.pdf' to generate these weights, ensuring that closer observations have a greater impact. The **'scipy.stats.norm.pdf'** function generates the probability density of the normal distribution, ensuring that the weight assigned to each observation reflects its proximity to the target age.
- The **'nadaraya_watson'** function then uses these weights to compute a weighted average of the observed weights, effectively smoothing the growth curve. This is achieved with the help of the in-built **'numpy.sum function'**, which sums the weighted values and normalizes the result, providing an estimate of the lamb's weight at any given age.
- Finally, the **'estimate_weight'** function applies the Nadaraya-Watson Estimator to individual lamb data, filtering the dataset using pandas functions like **'dropna'** to handle missing values and values to extract relevant data for analysis. This approach allows for the prediction of a lamb's weight at specific ages, even when direct measurements are not available, resulting in a continuous and smooth growth curve for detailed analysis.

Growth Curve with different Bandwidth

Figure 3.1 shows the application of the Nadaraya-Watson estimator to model the growth curve (for example, lamb no. 612), with varying bandwidths (h). The figure compares the estimated weight curves for different bandwidth values (10, 20, and 30), highlighting the effect of bandwidth selection on the smoothness and accuracy of the growth curve estimation.

Effect of Bandwidth on Estimation:

The bandwidth parameter in non-parametric regression controls the trade-off between bias and variance in the model. A smaller bandwidth (e.g., 10) leads to lower bias but higher variance, meaning the model will closely follow the data but might capture noise as well. Conversely, a larger bandwidth (e.g., 30) results in higher

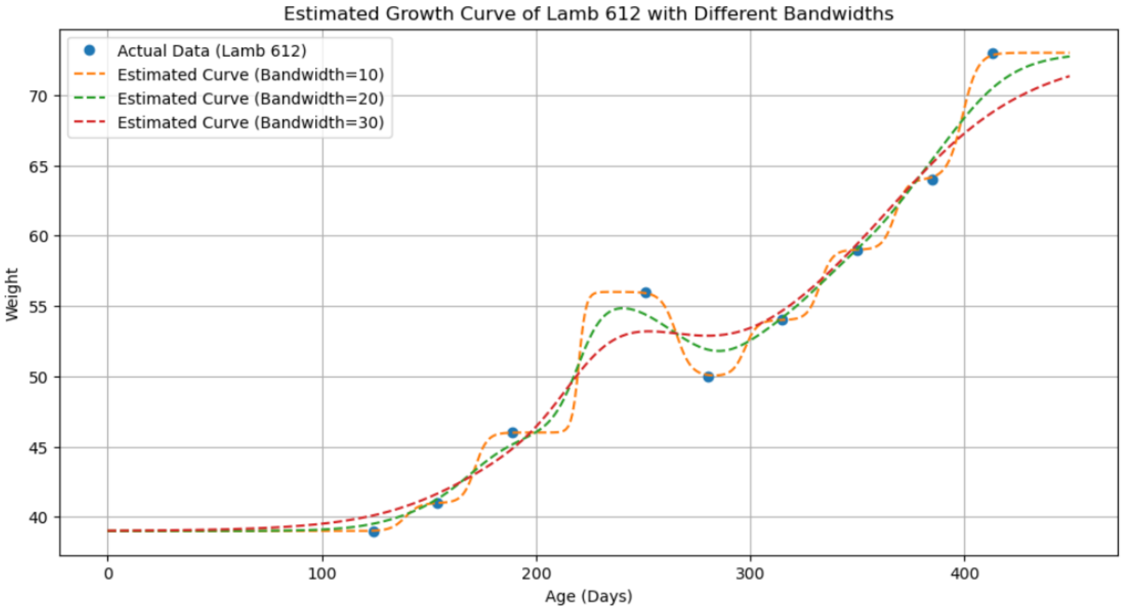


Figure 3.1: The figure depicts the estimated growth curve of lamb 612 using different bandwidths.

bias but lower variance, offering a smoother curve that generalizes better but might miss minute details of the data. Choosing an appropriate bandwidth is crucial for accurately estimating the underlying trend without overfitting or underfitting.

In conclusion, selecting the right bandwidth is essential in non-parametric regression, as it has a direct impact on the model's ability to capture the underlying trend accurately. A well-chosen bandwidth balances bias and variance, allowing the model to generalize effectively while avoiding the dangers of overfitting or underfitting. This careful balance ensures that the model not only fits the observed data closely but also remains robust and predictive when applied to new data, making it an important decision point in modeling tasks. The choice of bandwidth is a key factor in achieving reliable and interpretable growth curve predictions in any practical application.

3.1.4 Weighted Local Linear Fit Regression Technique

Weighted Local Linear Fit Regression is another non-parametric technique that extends the concept of kernel smoothing by fitting a linear model to localized subsets of the data. Instead of simply averaging the data points, this method fits a linear regression model within each local neighborhood, accounting for the slope of the data [9]. This approach provides better estimates in regions where the growth curve is not constant, particularly when the relationship between age and weight changes more rapidly.

This method is useful in longitudinal data analysis because it can adjust to local differences in growth patterns, resulting in more accurate estimates than simpler smoothing strategies. For the lamb dataset, weighted local linear fit regression captures minute changes in growth rate, enhancing weight prediction for unobserved ages.

Application in the Lamb Data

The technique works by fitting a local linear model to the data points within a neighborhood around the target age, rather than merely averaging the data points. The weights assigned to each observed data point are based on their proximity to the target age, with points closer to the target having a greater influence on the estimated value [10]. This is achieved through the use of a kernel function, which adjusts the influence of each data point depending on its distance from the target age. By considering both the value and the trend of the data in the local region, weighted local linear fit regression provides more accurate and responsive estimates, especially in regions where the growth curve exhibits non-linear patterns. This makes it a powerful tool for capturing subtle variations in lamb growth that might be missed by simpler smoothing methods.

Mathematical Formulation

Local linear regression is a method that fits a linear model to a subset of the data near the point of interest. This approach combines the flexibility of non-parametric methods with the interpretability of linear models.

For a point x (the specific age at which the weight prediction is being made), a linear model is fitted to the data points in its vicinity, with the inclusion of an error term ϵ_i to account for deviations between observed and predicted values. In this context:

$$y_i = \beta_0 + \beta_1(x_i - x) + \epsilon_i$$

- x_i represents the observed ages of the lambs at which weight measurements were recorded.
- y_i denotes the actual observed weights of the lambs corresponding to the ages x_i.
- β_0 is the intercept of the locally fitted linear model, which estimates the weight of the lamb at the specific age x.
- β_1 is the slope of the locally fitted linear model, indicating the rate of change in weight with respect to age in the local neighborhood of x.
- ϵ_i is the error term, capturing the difference between the observed weight y_i and the weight predicted by the linear model for age x_i.

Minimizing the Error

The goal is to minimize the weighted sum of squared errors, accounting for the error term ϵ_i:

$$\text{minimize} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(\epsilon_i)^2 = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(y_i - \beta_0 - \beta_1(x_i - x))^2$$

Here, $K(\cdot)$ is a kernel function that determines the weight of each data point based on its distance from x, and h is the bandwidth controlling the neighborhood size. The error term $\epsilon_i = y_i - \beta_0 - \beta_1(x_i - x)$ is squared and weighted by the kernel function.

Weighted Matrix and Vectors with Error Term

The weighted matrix $W(x)$ represents the weights assigned to each data point based on its distance from the target point x. The weights are determined using a kernel function, which influences how much each data point contributes to the local linear regression at x.

$$W(x) = \text{diag}\left(K\left(\frac{x_1 - x}{h}\right), K\left(\frac{x_2 - x}{h}\right), \ldots, K\left(\frac{x_n - x}{h}\right)\right)$$

Here, $W(x)$ is a diagonal matrix where each diagonal entry is the weight assigned to a corresponding data point. The weight is calculated using the kernel function applied to the normalized distance between each observed age x_i and the target age x.

The design matrix X contains the features (predictors) used to fit the linear model and the response vector Y contains the actual observed values that we are trying to model. In this case, it contains the weights of the lambs corresponding to the observed ages.

$$X = \begin{pmatrix} 1 & (x_1 - x) \\ 1 & (x_2 - x) \\ \vdots & \vdots \\ 1 & (x_n - x) \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Parameter Estimation with Error Term

The parameters $\hat{\beta}(x)$, including the error term, are estimated by minimizing the error sum of squares:

$$\hat{\beta}(x) = (X^T W(x) X)^{-1} X^T W(x) Y$$

where

$$\hat{\beta}(x) = \begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix}$$

The error term ϵ_i is implicitly handled through this optimization process by minimizing the weighted least squares error.

Prediction with Error Term

The predicted value at x still focuses on the intercept term of the locally fitted linear model:

$$\hat{y}(x) = \hat{\beta}_0(x) + \epsilon$$

Here, ϵ represents the error in the prediction. However, the fitted model inherently accounts for minimizing this error during the regression process.

The local linear regression method was implemented using Python with the following key components:

- **Gaussian Kernel Function**: The `gaussian_kernel` function calculates weights for each data point based on its proximity to the target age, ensuring nearby points have more influence.
- **Local Linear Regression**: The `local_linear_regression` function fits a linear model to data points near the target age by constructing a weight matrix W and solving the weighted least squares problem to estimate the coefficients $\hat{\beta}(x)$.
- **Weight Estimation**: The `estimate_weight_llr` function estimates a lamb's weight at a specific age by applying the local linear regression method to the relevant data.
- **Weight Curve Estimation**: The `estimate_weight_curve_llr` function generates a lamb's weight curve over a range of ages by estimating weights at each point.

After applying non-parametric models like the Nadaraya-Watson estimator and weighted local linear regression, the analysis was extended to parametric models. Non-parametric methods, while flexible, are limited by their reliance on local data and sensitivity to bandwidth selection. Parametric models, with their predetermined functional forms, provide more accurate interpretations and can generalize trends throughout the entire dataset. Exploring parametric models, such as polynomial or multiple regression, enables a more comprehensive understanding of lamb growth, helping to identify key influencing factors and develop robust predictive models that capture both local variations and global trends [11].

3.2 Parametric Models

Parametric models are essential in longitudinal data analysis as they offer a structured approach to modelling the relationship between variables over time using specific mathematical functions. These models assume a defined functional form, allowing for easy interpretation and prediction beyond observed data. Parametric models in longitudinal studies effectively capture the effects of many covariates and provide a clear understanding of how these factors influence the outcome across time [12]. By employing these models, such as **Polynomial Regression**, the analysis aims to make predictions and gain insights into the underlying trends and patterns in the data, allowing a more comprehensive understanding of the growth dynamics.

Unlike non-parametric models, which do not assume a predefined relationship between predictors and outcomes, parametric models like polynomial regression assume that specific relationships may exist between predictors and the outcome. For instance, in lamb growth analysis, such models can explore how factors like litter size or gender affect weight over time [14]. By fitting a polynomial function, parametric models can capture potential nonlinear relationships and interactions, providing a structured approach to studying how these factors affect development dynamics [13, 15].

3.2.1 Polynomial Regression Model

Polynomial regression is a versatile technique used in longitudinal data analysis to model complex relationships between time and the outcome variable by fitting a polynomial function to the data. This method extends simple linear regression by incorporating polynomial terms, which allows it to capture nonlinear trends and provides a more flexible approach to modeling growth or change over time [16]. In longitudinal studies, polynomial regression can effectively represent the progression of the outcome variable, accommodating varying growth rates and patterns.

Data Pre-processing

Before fitting the polynomial regression model, the data underwent pre-processing to ensure its suitability for analysis. The pre-processing involved three main steps:

1. **Handling Missing Data** : Missing values in the dataset were addressed by using the `dropna()` function. This function removes any rows with missing values in the specified columns, ensuring that only complete records are used for modeling:

   $$\text{cleaned\_data} = \text{final\_comb\_mulx.dropna()}$$

   Here `final_comb_mulx` is the dataframe created in the code.

2. **Adding the 'Litter_size' Column** : A new column referred to as `Litter_size` was created to represent the number of lambs born to the same mother on the same date. This was achieved by grouping the data by the `Dam` (mother) and `D.O.B` (date of birth) and counting the number of entries within each group.

   This was achieved using the following function:

   $$\text{df['Litter\_size']} = \text{df.groupby(['Dam', 'D.O.B'])['No.'].transform('count')}$$

   The function groups the data by `Dam` and `D.O.B` and counts the number of lambs within each group, thereby providing the litter size for each lamb. This approach ensures that the `Litter_size` feature accurately reflects the number of siblings for each lamb, which is crucial for understanding growth patterns and interactions in the polynomial regression model.

3. **Categorical Data Encoding** : Categorical columns, such as ‘Gender‘ and ‘Type‘, were converted into numerical values using one-hot encoding with the ‘pd.get_dummies()‘ function:

   ```
   pd.get_dummies(final_comb_mulx, columns=['Gender', 'Type'])
   ```

The pre-processing steps yield a refined dataset with complete records and an added Litter_size feature, essential for capturing the influence of litter size on lamb growth (Table 3.1). This final dataset was now ready for polynomial regression modeling.

Lamb No.	Age (Days)	Weight	Litter_size	Gender_Male	Type_pedigree
186	103	38.0	1.0	0	0
186	130	45.0	1.0	0	0
186	157	45.0	1.0	0	0
187	124	45.0	2.0	1	0
188	123	38.0	2.0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
2113	237	43.0	1.0	0	1
2113	270	46.0	1.0	0	1
2113	298	48.0	1.0	0	1
2113	326	51.0	1.0	0	1
2113	369	60.0	1.0	0	1

Table 3.1: Summary of the final version of data for modeling.

Mathematical Formulation

The general mathematical formulation for polynomial regression is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$

where:

- y_i is the outcome variable (e.g., lamb weight) at time x_i,
- x_i is the time or age variable,
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the polynomial terms,
- p is the degree of the polynomial,
- ϵ_i represents the error term, capturing the deviation of observed values from the fitted polynomial.

In context of the lamb dataset:

Independent Variables (Features): The matrix X contains the predictors:

$$X = \begin{pmatrix} \text{Age (Days)} \\ \text{Litter\_size} \\ \text{Gender\_M} \\ \text{Type\_pedigree} \end{pmatrix}$$

Dependent Variable (Target): The vector y contains the outcome variable:

$$y = \text{Weight}$$

Polynomial Transformation :

Polynomial transformation of the features is done using a third-degree polynomial:

$$X_{\text{poly}} = \begin{pmatrix} 1 & \text{Age (Days)} & (\text{Age (Days)})^2 & (\text{Age (Days)})^3 & \text{Litter\_size} & \text{Litter\_size} \cdot \text{Age (Days)} & \cdots \end{pmatrix}$$

This includes interactions between features and higher-order terms.

Linear Regression Model :

The polynomial regression model is represented as:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

where $x_{i,j}$ are the transformed features and β_j are the coefficients estimated by the model.

Model Evaluation

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Measures the average squared difference between observed and predicted values.

R-squared (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Indicates the proportion of variance in the dependent variable explained by the model.

In summary, the polynomial regression model developed here provides a robust framework for understanding the relationship between lamb weight and various predictor variables, including age, litter size, gender, and pedigree type. By using polynomial transformations and interaction terms, the model is capable of capturing the nonlinear growth patterns and complex interactions among the factors influencing lamb weight. The evaluation metrics, Mean Squared Error (MSE) and R-squared (R²), are employed to assess the model's performance in accurately predicting lamb weights. The detailed results and analysis of these models are presented in the following chapter, where the effectiveness of this approach are further explored.

Chapter 4

Results

This chapter provides a thorough evaluation of the results produced by the models discussed in the previous sections. Each model's performance is assessed using a variety of key metrics. These metrics are carefully selected to provide a detailed evaluation of the model's ability to predict lamb growth patterns with high accuracy.

4.1 Nadaraya-Watson Estimation

The Nadaraya-Watson estimator is a non-parametric regression method that was used to estimate the relationship between the age of lambs and their corresponding weights. This section explores the application of this technique to model the growth pattern of a specific lamb (Lamb 612). The results include an analysis of the estimated weight at a particular age compared to the actual weight, the overall growth curve, and an evaluation of the model's performance across different bandwidths using Mean Squared Error (MSE).

4.1.1 Bandwidth Selection and Model Performance

The choice of bandwidth in the Nadaraya-Watson estimator significantly affects the smoothness and accuracy of the estimated growth curve. To evaluate the model's performance, the Mean Squared Error (MSE) was calculated for different bandwidths using cross validation method.

Leave-One-Out Cross-Validation (LOO-CV)

Leave-One-Out Cross-Validation (LOO-CV) assesses the model's performance by testing it on one data point at a time. For each data point, the model is trained using all other data points, excluding the one being tested. The model then predicts the value of the excluded data point, and the error between the predicted and actual value is recorded. This process is repeated for every data point in the dataset. The errors are then averaged to provide a complete picture of the model's overall accuracy [17].

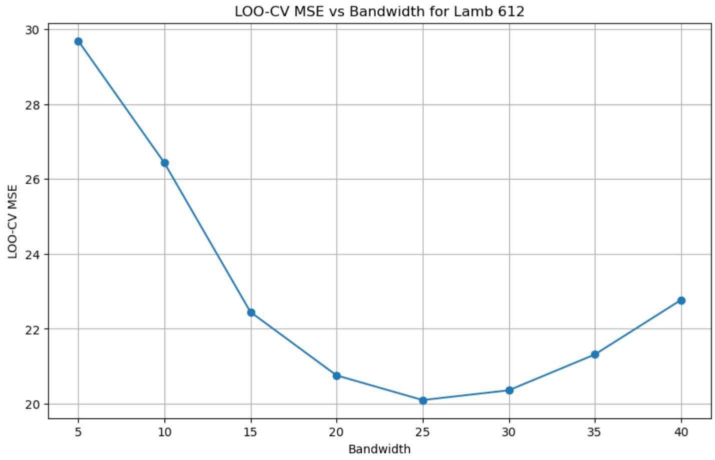


Figure 4.1: The plot for MSE was calculated using the cross validation method for different bandwidths.

Figure 4.1 shows the relationship between bandwidth LOO-CV and MSE for Lamb 612. As the bandwidth increases, the LOO-CV MSE initially decreases, reaching its lowest point around a bandwidth of 25 before rising again. This behavior suggests that a bandwidth of 25 provides the best balance between bias and variance for the model, minimizing the prediction error. Hence, the optimal bandwidth for estimating the growth curve of Lamb 612 based on this cross-validation analysis is 25.

4.1.2 Visualizing the Estimated Weight at a Specific Age

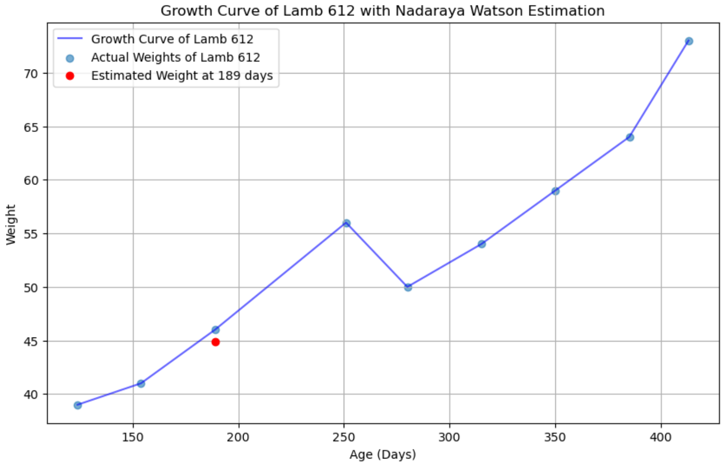


Figure 4.2: The figure shows the estimated weight of lamb 612 at the age of 189 days.

The estimated weight for Lamb 612 at the age of 189 days was found to be 44.86 kg, whereas the actual recorded weight at that age was 46 kg (see figure 4.2). This close prediction demonstrates the effectiveness of the Nadaraya-Watson estimator in capturing the weight. The minor difference between the estimated and actual weights reflects the model's precision, influenced by the selected bandwidth which was 25 in this case (see figure 3.1 for the estimated curve with different bandwidths).

4.1.3 Estimated Growth Curve Using Nadaraya-Watson

Figure 4.3 highlights that the estimated curve closely follows the actual data points, effectively modeling the overall growth pattern of the lamb. However, some deviations are observed, particularly in the middle range of ages where the curve slightly underestimates the weight. These variations could be due to bandwidth selection or inherent variability in the data.

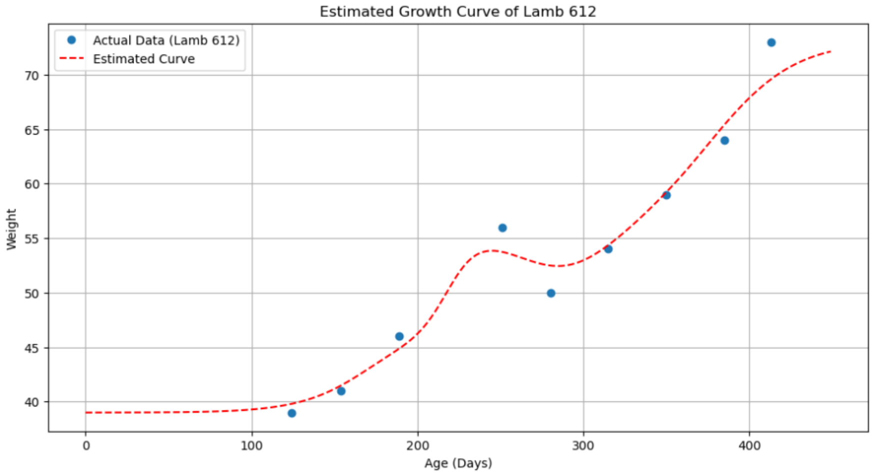


Figure 4.3: The figure depicts the estimated growth curve of lamb 612 with the bandwidth 25.

4.2 Weighted Local Linear Fit Regression

Weighted Local Linear Fit (LLR) Regression is another non-parametric method for estimating lamb growth curves. Unlike the Nadaraya-Watson estimator, LLR takes into account the local slope around each point, potentially providing a more accurate fit, particularly in areas where the growth rate changes. This section evaluates the application of LLR to model the growth of Lamb 612, presenting the estimated weight at a specific age and comparing the overall growth curves across different bandwidths. By adjusting the bandwidth, the flexibility of the curve can be controlled, allowing for a balance between overfitting and underfitting. The effectiveness of the LLR method is highlighted by its ability to capture small variations in the growth trajectory, making it a robust tool for modeling complex longitudinal data.

4.2.1 Optimal Bandwidth Selection for the Local Linear Regression Model

In non-parametric regression methods such as local linear regression, bandwidth selection is key to balancing bias and variance. A smaller bandwidth increases the model's sensitivity to data, risking overfitting by capturing noise rather than the true pattern. Conversely, a larger bandwidth smooths the model, reducing variance but potentially introducing bias. Thus, choosing an optimal bandwidth is critical for developing a model that generalizes effectively to new data.

Bandwidth Selection Using k-Fold Cross-Validation

To reduce the risk of overfitting, k-fold cross-validation was used as a more reliable method. The dataset was divided into k folds, with the model trained on k-1 folds and validated on the remaining fold. This process was repeated k times, and the errors were averaged to provide a stable performance estimate for each bandwidth. Cross-validation across various bandwidths identified the one that minimized the mean squared error (MSE), effectively balancing model complexity and generalization.

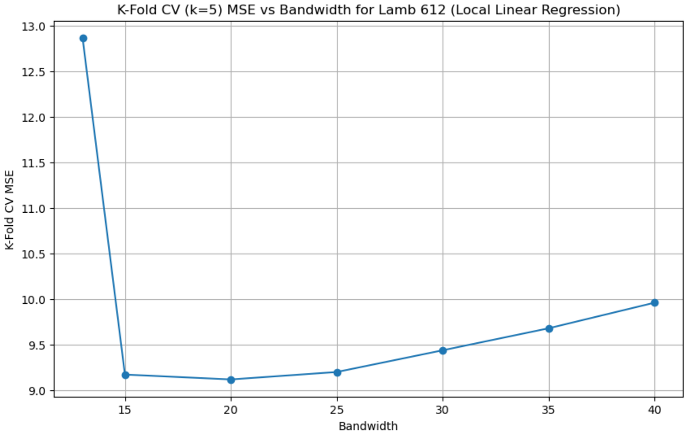


Figure 4.4: The plot for choosing optimal bandwidth using 5-fold cross validation is shown in the figure.

Figure 4.4 shows the relationship between the k-fold cross-validated mean squared error (MSE) and different bandwidth values for Lamb 612 using the Local Linear Regression (LLR) model. A range of bandwidths from 15 to 40 was evaluated using 5-fold cross-validation. The MSE initially decreases sharply as bandwidth increases from 15 to 20, reaching a minimum at 20. Beyond this point, the MSE begins to rise gradually, indicating that lower bandwidths can lead to overfitting, while larger bandwidths introduce bias.

The minimum MSE at a bandwidth of 20 suggests this is the optimal choice for balancing bias and variance in the LLR model for this dataset. This bandwidth achieves the best generalization performance, effectively minimizing the risk of overfitting while maintaining model accuracy.

4.2.2 Weight Estimation for a Specific Age Using LLR

Using the Local Linear Regression (LLR) method, the estimated weight for Lamb 612 at 189 days old was calculated to be 46.02 kg. This estimate is almost identical to the actual recorded weight of 46 kg for that same day (see figure 4.5). This close match between the estimated and actual weights highlights the accuracy of the LLR method in predicting the lamb's weight at 189 days. The small difference of only 0.02 kg shows that the LLR method is highly effective for this specific prediction, suggesting it can reliably capture the lamb's growth pattern.

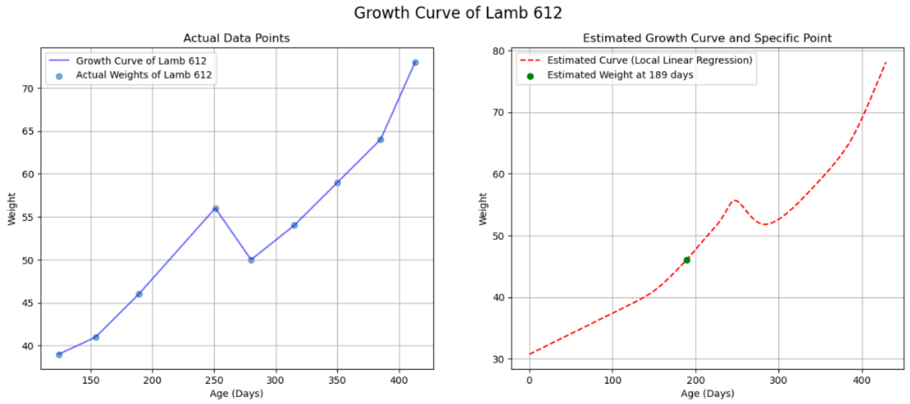


Figure 4.5: Estimated growth patterns of lamb 612 using a bandwidth of 20 with the Local Linear Regression (LLR) method. The left figure shows the actual growth curve from observed data just by joining the actual data points, while the right figure presents the smoothed growth curve using LLR.

4.2.3 Growth Curve Analysis with Different Bandwidths

Figure 4.6 illustrates the estimated growth curves for Lamb 612 using the Local Linear Regression (LLR) method with varying bandwidths (20, 30, and 40). The graph includes:

- **Black Dots:** Representing the actual observed weights of Lamb 612 at various ages.
- **Red Dashed Line:** The estimated growth curve with a bandwidth of 20.
- **Green Dashed Line:** The estimated growth curve with a bandwidth of 30.
- **Blue Dashed Line:** The estimated growth curve with a bandwidth of 40.

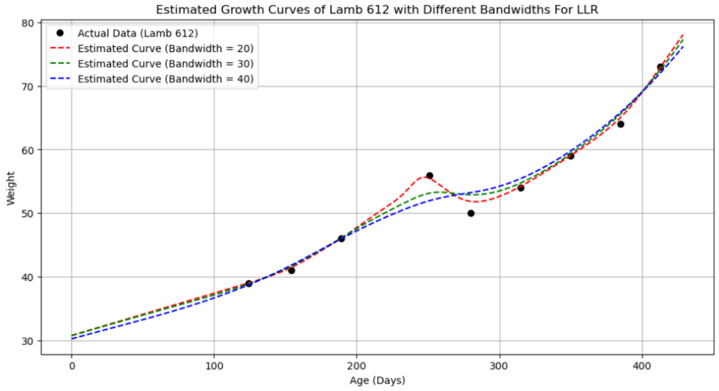


Figure 4.6: Figure highlights the estimated growth curves of lamb 612 with different bandwidths.

The different curves provide insight into how the choice of bandwidth affects the smoothness and accuracy of the model. The curve with a bandwidth of 20 is more responsive to fluctuations in the data, while the curves with larger bandwidths (30 and 40) are smoother and may underfit in some areas.

4.3 Polynomial Regression Model

This section focuses on the implementation and evaluation of a polynomial regression model to analyze the growth trends of lambs. The model evaluates how various factors, such as gender, litter size, and pedigree type, influence lamb weight over time. By applying a third-degree polynomial transformation, the model aims to capture the relationships between these variables and provide a deeper understanding of the general growth patterns observed in lambs. The results also provide information about the predictive accuracy of the model.

4.3.1 Model Implementation and Evaluation

Data Splitting and Polynomial Transformation

The dataset was first cleaned by removing any rows with missing values. The independent variables selected for the model were age (days), litter size, gender, and breed, with Weight as the dependent variable. The data was then split into training and testing sets using a 75%-25% ratio to ensure robust model evaluation. A third-degree polynomial transformation was applied to the features to capture potential nonlinear relationships between the predictors and the outcome variable.

Model Fitting and Prediction

A linear regression model was fitted using transformed polynomial features, specifically applying a third-degree polynomial function to capture the growth curve pattern observed in the lambs. The model was trained on the training set and then used to predict the weights in the test set.

Model Evaluation and Residuals

The model's performance was evaluated using the Mean Squared Error (MSE) and R-squared (R^2) metrics. The MSE was turned out to be 33.24, indicating the average squared difference between observed and predicted values. The R^2 value of 0.66 suggests that approximately 66% of the variance in the lamb's weight can be explained by the model.

Coefficient Analysis & Inference

The model's coefficients, as shown in the table 4.1, represent the contribution of each feature and its polynomial interactions to the prediction of lamb weight. The table also lists the actual and predicted weights for the test data, along with the residuals, allowing for comparison between the model's predictions and the true values.

The results of the polynomial regression model indicates that the age positively affects weight, while gender, litter size, and pedigree show varying effects depending on their interactions. Notably, the impact of litter size and pedigree type is non-linear, indicating that these factors interact in more complex ways

Feature	Coefficient
Intercept	-1.817032e-08
Age (Days)	6.065441e-01
Litter_size	1.161442e-02
Gender_M	-6.114325e-02
Type_pedigree	-2.194126e-01
Age (Days)^2	1.019946e-03
Age (Days) * Litter_size	1.643994e-03
Age (Days) * Gender_M	-7.733598e-02
Age (Days) * Type_pedigree	2.288406e-02
Litter_size^2	3.161432e-02
Litter_size * Type_pedigree	4.970344e+00
Gender_M * Type_pedigree	9.042818e-02
Gender_M * Type_pedigree^2	-7.198412e-01
Type_pedigree^2	2.794738e+00
Age (Days)^2 * Litter_size	8.267225e-06
Age (Days)^2 * Gender_M	2.734427e-04
Age (Days)2 * Type_pedigree	-3.950446e-04
Litter_size^2 * Gender_M	-1.953407e-03
Litter_size^2 * Type_pedigree	2.934839e-01
Gender_M^2 * Type_pedigree	1.888253e-02
Age (Days)^3	-9.873096e-08
Age (Days)^2 * Litter_size	2.739828e-04
Age (Days)^2 * Gender_M	-9.703310e-02
Age (Days) * Litter_size^2	-9.138589e-03
Age (Days) * Litter_size * Gender_M	-1.345224e-02
Age (Days) * Litter_size * Type_pedigree	3.791578e-02
Litter_size^3	1.807500e-04
Litter_size^2 * Gender_M	2.645430e-02
Litter_size * Gender_M^2	-1.791312e-01
Litter_size2 * Type_pedigree	-3.493048e-02
Gender_M^2 * Type_pedigree^2	1.791438e+00
Gender_M * Type_pedigree^3	-2.194126e-01

Actual	Predicted	Residual
50.0	50.025592	-0.025592
21.0	36.522120	-15.522120
55.0	54.223884	0.776116
36.0	38.490177	-2.490177
44.0	37.153738	6.846262
⋮	⋮	⋮
56.0	54.392546	1.607454
42.0	42.073278	-0.073278
36.0	37.153738	-1.153738
32.0	32.957940	-0.957940
38.0	39.676974	-1.676974

Table 4.1: The table describes the result of the polynomial regression model.

than simple linear relationships would suggest. Overall, the model captures the nuanced influences on lamb growth effectively.

The polynomial regression model for the lamb data demonstrates complex relationships between growth and various factors. The coefficient for Age (Days) (0.606) indicates a positive relationship, with lambs gaining weight as they age. The quadratic term for age (0.00102) shows that this weight gain accelerates over time. Litter_size has a small positive effect (0.0116), with its interaction with age (0.00164) slightly enhancing growth. The coefficient for Gender_M (-0.0611) indicates that male lambs weigh more compared to females, aligning with the dataset's observations. The type of lamb (-0.2194 for pedigree) affects growth, with pedigree lambs generally weighing less. Complex interactions, such as Age (Days) × Gender (-0.0773) and Age (Days) × Type_pedigree (0.0229), reveal that growth patterns are moderated by these factors, suggesting that while growth dynamics are influenced by a combination of age, gender, and type, the polynomial and interaction terms effectively capture the nuanced effects observed in the data.

4.3.2 Weight Predictions and Growth Curves

In this section, the focus will be on analyzing the predicted weight and growth curves for individual lambs based on the polynomial regression model. By plotting these estimated growth curves, it becomes possible to generalize the growth trends for lambs, taking into account various influencing factors such as age, gender, litter size, and breed. These insights will help to better understand how these factors contribute to the growth patterns observed in lambs over time.

Growth Curve Prediction for Lambs

Lamb 612 is a female with a litter size of 2 and is of type-pedigree. With the help of the model growth curve, the lamb was produced (see figure 4.7). The close alignment between the actual data points and the predicted curve suggests that the model is effective in capturing the growth pattern of Lamb 612, demonstrating its usefulness in forecasting weight trends.

The expected growth curve and the actual weight data for four lambs (Lamb 711, Lamb 825, Lamb 829, and Lamb 931) are displayed in figure 4.8, all of which share the same characteristics: they are female, have a litter size of two, and are of the pedigree type. The red dashed line represents the predicted growth curve generated by the regression model, which generally follows the upward trend of the actual data points, indicating the model's effectiveness in capturing the overall growth pattern. Similarly, growth curves for lambs with different characteristics, such as varying litter sizes, genders, or pedigree types, can be produced using the same model, allowing for customized predictions across a diverse set of conditions.

4.3.3 General Trend in Growth Curves

In this section, the focus is on analyzing the general trends in growth curves for lambs based on various factors such as litter size, gender, and pedigree type. By using the developed polynomial regression model, predicted growth patterns are visualized, offering insights into how these factors influence the weight progression of lambs over time. The analysis will explore how different combinations of these variables affect

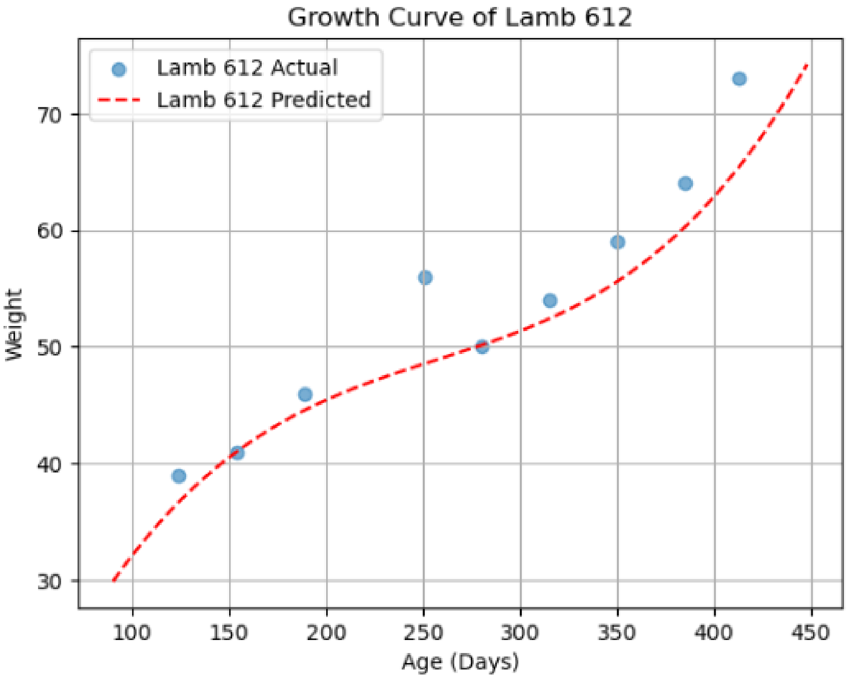


Figure 4.7: Figure represents the predicted growth curve of lamb 612 by using the model's estimated weights.

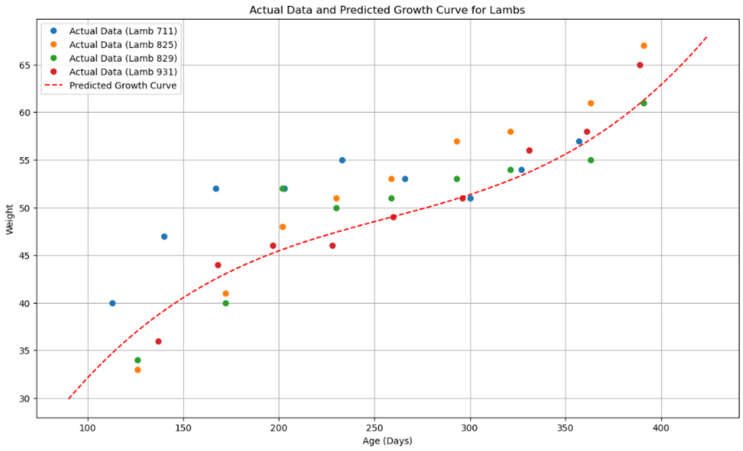


Figure 4.8: Figure shows the general predicted growth curve with certain characteristics compared with actual weights of four lambs.

the overall growth trajectory, helping to understand the diverse growth characteristics observed in commercial and pedigree lambs.

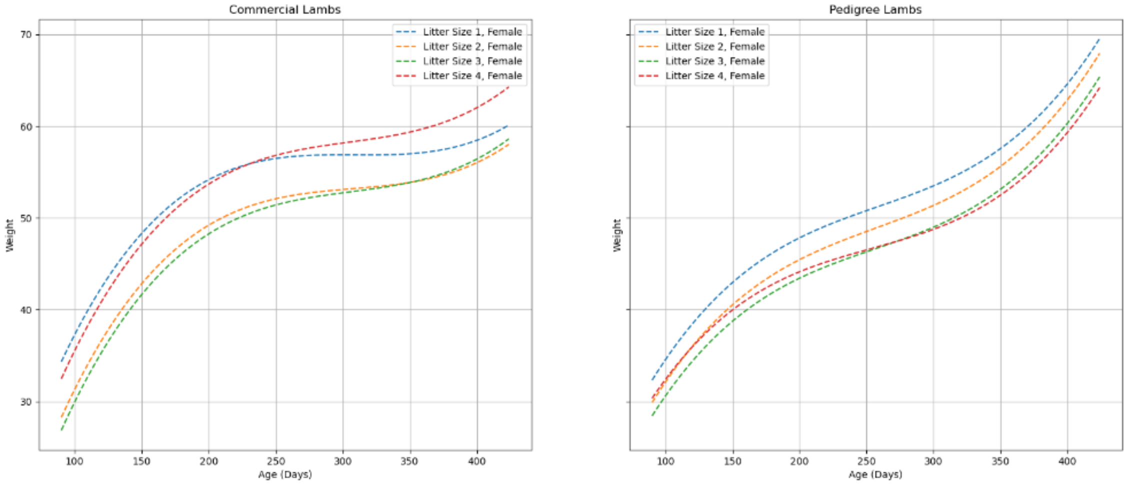


Figure 4.9: General trend of female lamb growth curves for pedigree and commercial breeds with different litter sizes.

Figure 4.9 highlights distinct growth patterns between commercial and pedigree lambs based on age and litter size. For commercial lambs, the growth curve shows a rapid initial weight gain, indicating that these lambs put on weight quickly in their early stages. In contrast, pedigree lambs exhibit a more gradual and steady increase in weight over time, which eventually results in higher weights compared to commercial lambs as they mature.

Examining the effect of litter size, the graph reveals significant trends, particularly in pedigree lambs. As the litter size increases, the growth curves shift downward, indicating lower weight gains. This trend is clear for pedigree lambs, where those from larger litters (especially litter size 4) initially show higher weights but then experience slower growth compared to those from smaller litters. Interestingly, commercial lambs with a litter size of 4 deviate from this pattern, showing unexpectedly high weight gains. However, this is likely due to a small sample size for this group, which introduces a bias and affects the general trend in the growth curve for commercial lambs.

Figure 4.10 presents the growth curves of male lambs, differentiated by litter size, for both commercial and pedigree types. In the case of commercial lambs, the growth pattern is relatively uniform across different litter sizes, with larger litter sizes generally correlating with slightly lower weights. Interestingly, lambs from larger litters (e.g., litter size 4) tend to catch up in weight over time, though not surpassing those from smaller litters.

Litter size has a stronger influence on pedigree lambs. As litter size increases, there is a clear downward shift in the growth curve, indicating that lambs from larger litters tend to have slower growth and lower weights overall. Lambs from litter size 1 consistently show the highest weights, while those from litter size 4 show significantly lower weights, with a noticeable gap that remains and even widens over time. This

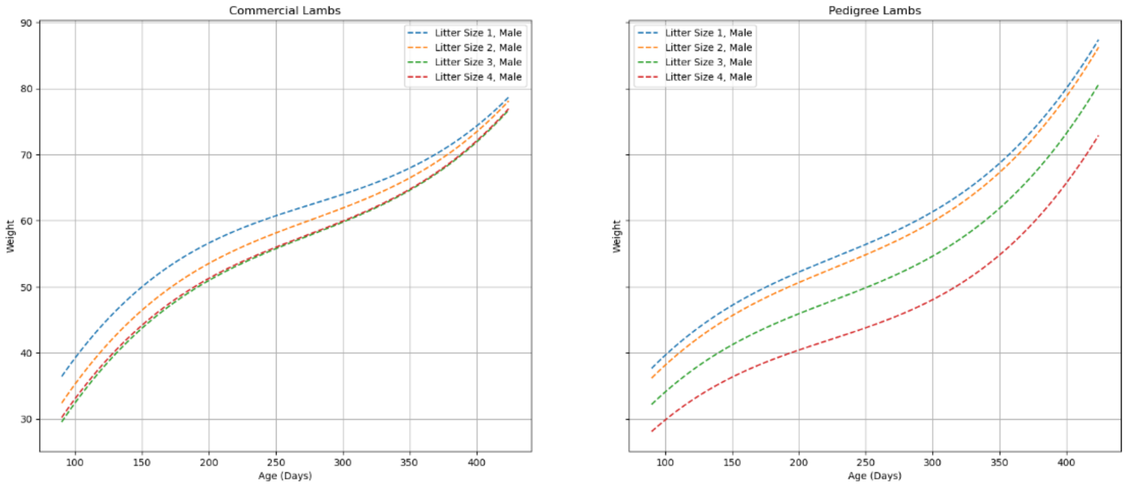


Figure 4.10: General trend of male lamb growth curves for pedigree and commercial breeds with different litter sizes.

suggests a significant negative correlation between litter size and growth rate in pedigree lambs.

The analysis of general growth trends in both male and female lambs reveals key differences based on litter size and breed. Commercial lambs, regardless of gender, tend to have more uniform growth patterns with minor deviations, while pedigree lambs exhibit a clearer inverse relationship between litter size and weight gain. These patterns highlight how factors distinctly influence growth trajectories across different groups.

Chapter 5

Discussion

This chapter thoroughly evaluates three modeling approaches to predicting growth curves. The models are compared in terms of how well they capture and generalize growth patterns from longitudinal data. Each method's key advantages and disadvantages are summarized, along with their effectiveness in achieving accurate growth curve estimation.

5.1 Model Comparison

Criteria	**Nadaraya-Watson Estimation**	**Weighted Local Linear Fit Regression**	**Polynomial Regression**
Results	Optimal bandwidth = 25	Optimal bandwidth = 20	Polynomial Degree = 3, 66% accuracy
Model Dependence on Data	Heavily relies on existing data points for estimation.	Strongly dependent on existing data points for local fitting.	Generalizes growth trends using a specific functional form.
Performance Range	Best within the data range.	Performs well within the data range.	Capable of generalizing and predicting growth for new lambs.
Extrapolation Capability	Limited; struggles outside the data range.	Limited; struggles outside the data range.	Capable of extrapolating and applying the trend to new cases.
Parameter Selection	Bandwidth selection is critical, chosen via LOO-CV.	Bandwidth is key, optimized using k-fold cross-validation.	Degree of polynomial is crucial; balance needed to avoid overfitting or underfitting.
Strengths	Flexible and adapts to the shape of data. Effective at capturing complex patterns within the data.	Captures local trends accurately. Less sensitive to boundary effects.	Provides a clear and interpretable functional form. Capable of generalizing to new data.
Limitations	Sensitive to outliers. Limited extrapolation ability.	Computationally intensive for large datasets. Sensitive to kernel choice.	Risk of overfitting with higher-degree polynomials. May not capture local variability.

Table 5.1: Comparison of Nadaraya-Watson Estimation, Weighted Local Linear Fit Regression, and Polynomial Regression

Table 5.1 compares the three techniques: Nadaraya-Watson estimation, weighted local linear fit, and polynomial regression, focusing on their optimal parameters, data dependencies, and performance features. Nadaraya-Watson and Weighted Local Linear Fit are adaptable and successful within the data range but suffer with extension, whereas polynomial regression generalizes well and can extend, but with the risk of overfitting. Each method's strengths and limitations are determined by its dependency on bandwidth or polynomial degree, which affects adaptability and accuracy.

5.2 Observations

- **Gender-Based Weight Differences:** Male lambs consistently achieved higher weights from early ages compared to female lambs.
- **Weight Differences by Breed:** Commercial lambs, regardless of gender, typically have higher weights than pedigree lambs. This difference often led to commercial lambs being sold primarily for meat, while pedigree lambs were typically reserved for breeding purposes.
- **Growth Patterns:** The growth trajectories of commercial and pedigree lambs differed significantly, with commercial lambs having higher initial weights and continuously increasing thereafter. Pedigree lambs have lower weights in their early ages but grow faster later on. Regardless of breed, both types of lambs' growth rates slowed between 200 and 250 days of age before growing again.
- **Impact of Litter Size on Male Lambs:** The growth curve of male lambs showed a noticeable downward shift in trajectory as litter size increased, particularly in pedigree lambs. This indicates that lambs from larger litters are born with lower weights and maintain this lower weight pattern throughout their growth. In commercial lambs, the trajectory moves downward with bigger litter sizes, but the weight differential at later stages remains less evident. Litter sizes 3 and 4 almost have the same growth trend, with minor variations.
- **Impact of Litter Size on Female Lambs:** Female lambs also displayed a similar pattern where litter size affected growth trajectory. However, the impact was less significant in pedigree lambs with litter sizes of 3 and 4 compared to those from smaller litters. It also shows that the initial weights of litter size 4 were higher but thereafter fell when compared to smaller litters, indicating a general pattern of lower weights as litter size increases. In commercial lambs, except for the data for litter size 4, the growth curves showed significant differences in growth trajectories, highlighting the impact of litter size on weight development.

5.3 Conclusion

The analysis demonstrated the effectiveness of non-parametric methods in providing accurate estimates of lamb weights. These methods proved particularly flexible, allowing for the modeling of growth curves without the need for predefined assumptions about the relationship between age and weight. However, while highly effective for lambs with similar characteristics, the accuracy of these methods may decrease when applied to lambs with varying traits, as their performance is closely tied to the specific data used.

The study provides valuable insights into optimizing farm management practices by enabling precise prediction of lamb growth trajectories. By leveraging these predictive tools, farmers can enhance breeding strategies and efficiently allocate resources to maximize meat production, ultimately improving overall farm productivity.

5.3.1 Optimizing Breeding Strategies

Predictive Capabilities

These models allow for the prediction of growth patterns based on variables like gender, litter size, and breed type. By understanding how different genetic and environmental factors influence growth, farmers can make informed decisions about which traits to prioritize in their breeding programs. For instance, if a certain breed or litter size consistently results in higher weight gains at specific ages, breeding strategies can be adjusted to favor these characteristics.

Timing of Breeding

The ability to predict when lambs will reach optimal market weight can help in planning breeding cycles. For example, by using these models, farmers can time breeding to ensure that lambs reach peak growth during periods of high market demand or favorable environmental conditions.

5.3.2 Maximizing Meat Production

Growth Monitoring

The models provide detailed growth curves that help in monitoring lamb's weight gain over time. Farmers can identify and address growth issues early, such as undernutrition or disease, that could hinder the lamb's development. This ensures that lambs reach their full potential in terms of weight, directly impacting meat yield.

Resource Allocation

By predicting growth trajectories, farmers can allocate resources, such as feed and animal health care, more effectively. For example, lambs identified as slower-growing might be given additional resources to boost their growth, while faster-growing lambs might be managed differently to maximize overall farm efficiency.

These predictions help identify lambs best suited for specific purposes, such as meat production or breeding, and the resulting trend analyses provide valuable insights for optimizing management practices. Overall, these approaches offer practical tools for enhancing decision-making and improving outcomes in lamb management.

5.4 Future Work

The existing data could be used to conduct a more detailed temporal analysis of growth patterns, focusing on specific age intervals. This could help identify critical growth periods where management interventions (such as selective breeding or targeted care) might be most effective in influencing outcomes.

Utilizing the available data, cluster analysis could be conducted to group lambs with similar growth patterns. This could help identify distinct growth profiles within the population, which could then be used to develop more customized management strategies for different groups.

A long-term trend analysis could be performed to examine how growth patterns have changed over time. This could provide insights into whether certain factors, like breed or litter size, are becoming more or less influential over time and help predict future trends in lamb growth.

Further research could explore the use of more advanced machine learning methods, like deep learning or ensemble approaches, to enhance the predictive accuracy of lamb growth models. These techniques might reveal complex relationships between variables that were not captured.

Appendix A

Python Code

Section 2.2.2 - Data Transformation

```
# Extracting the columns that have dates
date_columns = df_2018testtr.columns[4:-2]  # Skip ’No.’, ’Sex’, ’D.O.B
    ’, ’Dam’, ’Litter_size’, and ’type’

lambs_list = []
age_list = []
weight_list = []
litter_size_list = []
gender_list = []
type_list = []

# Calculating the Age (Days)
for col in date_columns:
    age_days = (pd.to_datetime(col, format=’%d/%m/%Y’) - df_2018testtr[’
        D.O.B’]).dt.days
    lambs_list.extend(df_2018testtr[’No.’])
    age_list.extend(age_days)
    weight_list.extend(df_2018testtr[col])
    litter_size_list.extend(df_2018testtr[’Litter_size’])
    gender_list.extend(df_2018testtr[’Sex’])
    type_list.extend(df_2018testtr[’type’])

# Creating the final DataFrame
final_df2 = pd.DataFrame({
    ’No.’: lambs_list,
    'Age (Days)': age_list,
    'Weight': weight_list,
    'Litter_size': litter_size_list,
    'Gender': gender_list,
    'Type': type_list
})

# Sorting and reseting the index
final_df2.sort_values(by=['No.', 'Age (Days)'], inplace=True)
final_df2.reset_index(drop=True, inplace=True)

combined_df = pd.concat([final_df2, final_df2_2016,final_df2_2017,
   final_df2_2019,final_df2_2020,final_df2_2021,final_df2_2022],
   ignore_index=True)

# Sort and reset index
combined_df.sort_values(by=['No.', 'Age (Days)'], inplace=True)
combined_df.reset_index(drop=True, inplace=True)
```

Figure 2.8 - Growth Curve Of Lambs for the year 2018

```
# Plotting the Curve
plt.figure(figsize=(20, 20))

for lamb_no in final_df2['No.'].unique():
    lamb_data = final_df2[final_df2['No.'] == lamb_no]
    plt.plot(lamb_data['Age (Days)'], lamb_data['Weight'], marker='o',
       label=f'Lamb {lamb_no}')

plt.title('Growth Curve of Lambs', fontsize=24)
plt.xlabel('Age (Days)', fontsize=20)
plt.ylabel('Weight', fontsize=20)
#plt.legend()
plt.grid(True)
plt.show()
```

Figure 2.9 - Growth Curve comparison of lamb twins

```
lambs_to_compare = [453,454]
```

```
plt.figure(figsize=(10, 6))

for lamb_no in lambs_to_compare:
    lamb_data = combined_df[combined_df['No.'] == lamb_no].dropna()
    plt.plot(lamb_data['Age (Days)'], lamb_data['Weight'], marker='o',
       label=f'Lamb {lamb_no}')

plt.title('Growth Curve Comparison of Lamb Twins', fontsize=18)
plt.xlabel('Age (Days)', fontsize=14)
plt.ylabel('Weight', fontsize=14)
plt.legend(title='Lamb No.')
plt.grid(True)
plt.show()
```

Figure 2.10 - Growth Curve comparison of lambs based on litter size

```
lambs_to_compare = [839,842]
plt.figure(figsize=(10, 6))

for lamb_no in lambs_to_compare:
    lamb_data = combined_df[combined_df['No.'] == lamb_no].dropna()
    plt.plot(lamb_data['Age (Days)'], lamb_data['Weight'], marker='o',
       label=f'Lamb {lamb_no}')

plt.title('Growth Curve Comparison of Lambs based on Litter Size',
   fontsize=18)
plt.xlabel('Age (Days)', fontsize=14)
plt.ylabel('Weight', fontsize=14)
plt.legend(title='Lamb No.')
plt.grid(True)
plt.show()
```

Section 3.1.3 - Nadaraya-Watson Estimation Technique

```
# Gaussian Smoothing
def gaussian_kernel(distance, bandwidth):
    return norm.pdf(distance / bandwidth)

# Assigning Weights to data points
```

```
def nadaraya_watson(x, X, Y, bandwidth):
    weights = gaussian_kernel(x - X, bandwidth)
    return np.sum(weights * Y) / np.sum(weights)

def estimate_weight(lamb_no, target_age, bandwidth):
    lamb_data = combined_df[combined_df[’No.’] == lamb_no].dropna(subset
        =[’Weight’])
    X = lamb_data[’Age (Days)’].values
    Y = lamb_data[’Weight’].values
    estimated_weight = nadaraya_watson(target_age, X, Y, bandwidth)
    return estimated_weight

# Example: Estimating the weight of lamb 611 at age 200 days
lamb_no = 611
target_age = 200
bandwidth = 10  # Bandwidth Adjustment

estimated_weight = estimate_weight(lamb_no, target_age, bandwidth)
print(f”Estimated weight of lamb {lamb_no} at age {target_age} days: {
   estimated_weight}”)

Result : Estimated weight of lamb 611 at age 200 days: 48.9997207175899
```

Nadaraya-Watson Bandwidth Estimation Using LOO-CV (Figure - 4.1)

```
# Function to perform LOO-CV for a given lamb and bandwidth
def loo_cv(lamb_no, bandwidth):
    lamb_data = final_df2[final_df2[’No.’] == lamb_no].dropna(subset=[’
        Weight’])
    X = lamb_data[’Age (Days)’].values
    Y = lamb_data[’Weight’].values

    errors = []

    # Leave-One-Out Cross-Validation
    for i in range(len(X)):
        # Use all data except the i-th data point
        X_train = np.delete(X, i)
        Y_train = np.delete(Y, i)
        # Test on the i-th data point
        x_test = X[i]
        y_test = Y[i]

        # Estimate the weight using the Nadaraya-Watson method
        y_pred = nadaraya_watson(x_test, X_train, Y_train, bandwidth)

        # Calculate the squared error
        errors.append((y_test - y_pred) ** 2)

    return np.mean(errors)

# Define the lamb number
lamb_no = 612

# Define the bandwidths to test
bandwidths = [5, 10, 15, 20, 25, 30, 35, 40]

# Storing LOO-CV MSE for each bandwidth
loo_mse_values = []

for bandwidth in bandwidths:
    mse = loo_cv(lamb_no, bandwidth)
    loo_mse_values.append(mse)

# Plot LOO-CV MSE vs Bandwidth
plt.figure(figsize=(10, 6))
plt.plot(bandwidths, loo_mse_values, marker='o')
plt.title('LOO-CV MSE vs Bandwidth for Lamb 612')
plt.xlabel('Bandwidth')
plt.ylabel('LOO-CV MSE')
plt.grid(True)
plt.show()

# Printing the optimal bandwidth
optimal_bandwidth = bandwidths[np.argmin(loo_mse_values)]
print(f'Optimal Bandwidth based on LOO-CV: {optimal_bandwidth}')
```

```
Result:Optimal Bandwidth based on LOO-CV: 25
```

Figure 3.1

```
# Estimating the weight curve for lamb
def estimate_weight_curve(lamb_no, X, bandwidth):
    estimated_weights = []
    for age in X:
        estimated_weight = estimate_weight(lamb_no, age, bandwidth)
        estimated_weights.append(estimated_weight)
    return np.array(estimated_weights)

# Define the lamb number and the age range for estimation
lamb_no = 612
X = np.arange(0, 450, 1)  # Estimating weights from day 0 to 450
bandwidths = [10, 20, 30]  # Different bandwidths to compare

# Plot the actual data
plt.figure(figsize=(12, 6))
lamb_data = final_df2[final_df2['No.'] == lamb_no]
plt.plot(lamb_data['Age (Days)'], lamb_data['Weight'], 'o', label=f'
   Actual Data (Lamb {lamb_no})')

# Plot the estimated curves for different bandwidths
for bw in bandwidths:
    estimated_weights = estimate_weight_curve(lamb_no, X, bw)
    plt.plot(X, estimated_weights, label=f'Estimated Curve (Bandwidth={
       bw})', linestyle='--')

# Adding titles and labels
plt.title('Estimated Growth Curve of Lamb 612 with Different Bandwidths
   ')
plt.xlabel('Age (Days)')
plt.ylabel('Weight')
plt.legend()
plt.grid(True)
plt.show()
```

Weighted Local Linear Fit Regression Technique - Section 3.1.4

```
# Function to compute Gaussian kernel
def gaussian_kernel(distance, bandwidth):
    return norm.pdf(distance / bandwidth)

# Function to compute local linear regression estimate
def local_linear_regression(x, X, Y, bandwidth):
    n = len(X)
    weights = gaussian_kernel(X - x, bandwidth)

    # Construct weight matrix
    W = np.diag(weights)

    # Construct design matrix
    X_mat = np.vstack((np.ones(n), X - x)).T

    # Compute (X^T W X)^-1 X^T W Y
    beta = np.linalg.inv(X_mat.T @ W @ X_mat) @ (X_mat.T @ W @ Y)

    # Predicted value at x
    y_pred = beta[0]

    return y_pred

# Function to estimate weight for a given lamb and target age using
   local linear regression
def estimate_weight_llr(lamb_no, target_age, bandwidth):
    lamb_data = final_df2[final_df2['No.'] == lamb_no].dropna(subset=['
       Weight'])
    X = lamb_data['Age (Days)'].values
    Y = lamb_data['Weight'].values
    estimated_weight = local_linear_regression(target_age, X, Y,
       bandwidth)
    return estimated_weight

# Estimating the weight curve for lamb using local linear regression
def estimate_weight_curve_llr(lamb_no, X, bandwidth):
    estimated_weights = []
    for age in X:
        estimated_weight = estimate_weight_llr(lamb_no, age, bandwidth)
        estimated_weights.append(estimated_weight)
    return np.array(estimated_weights)

# Define the lamb number and the age range for estimation
lamb_no = 612
X = np.arange(0, 430, 1)  # Estimating weights from day 0 to 300
bandwidth = 15  # Bandwidth Adjustment

# Estimating the weight curve using local linear regression
estimated_weights_llr = estimate_weight_curve_llr(lamb_no, X, bandwidth)
```

Polynomial Regression Model - Section 3.2.1

```
#Data Pre-Processing
final_comb_mulx = pd.get_dummies(comb_mul_df, columns=['Gender','Type'],
    drop_first=True)
final_comb_mulx.dropna(subset=['Weight'], inplace=True)
final_comb_mulx.reset_index(drop=True, inplace=True)

final_comb_mulx=final_comb_mulx.dropna()
# Defining the independent variables (features) and dependent variable (
   target)
X = final_comb_mulx[['Age (Days)', 'Litter_size', 'Gender_M','
   Type_pedigree']]
y = final_comb_mulx['Weight']

# Spliting the data into training and testing sets (Cross Validation)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
   =0.25, random_state=42)

# Polynomial transformation (3rd order)
poly = PolynomialFeatures(degree=3)
X_poly_train = poly.fit_transform(X_train)
X_poly_test = poly.transform(X_test)

# Initializing and fitting the model
model = LinearRegression()
model.fit(X_poly_train, y_train)
```

```
# Predicting the weights for the test set
y_pred = model.predict(X_poly_test)

# Calculating the residuals
residuals = y_test - y_pred

# Evaluating the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f’Mean Squared Error: {mse}’)
print(f’R-squared: {r2}’)

Result: Mean Squared Error: 33.24284844649177
R-squared: 0.6604974506900636
```

Figure 4.3

```
# Define the lamb number and the age range for estimation
lamb_no = 612
X = np.arange(0, 450, 1)  # Estimating weights from day 0 to 300
bandwidth = 25  # Bandwidth Adjustment

estimated_weights = estimate_weight_curve(lamb_no, X, bandwidth)

# Plotting the actual data and the estimated curve
plt.figure(figsize=(12, 6))
lamb_data = final_df2[final_df2[’No.’] == lamb_no]
plt.plot(lamb_data[’Age (Days)’], lamb_data[’Weight’], ’o’, label=f’
    Actual Data (Lamb {lamb_no})’)
plt.plot(X, estimated_weights, label=’Estimated Curve’, color=’red’,
    linestyle=’--’)

plt.title(’Estimated Growth Curve of Lamb 612’)
plt.xlabel(’Age (Days)’)
plt.ylabel(’Weight’)
plt.legend()
plt.grid(True)
plt.show()
```

Optimal Bandwidth Section for LLR- 4.2.1,Figure 4.4

```
def k_fold_cv_llr(lamb_no, bandwidth, k=5):
    lamb_data = final_df2[final_df2[’No.’] == lamb_no].dropna(subset=[’
        Weight’])
    X = lamb_data[’Age (Days)’].values
    Y = lamb_data[’Weight’].values

    kf = KFold(n_splits=k, shuffle=True, random_state=1)
    mse_values = []

    for train_index, test_index in kf.split(X):
        X_train, X_test = X[train_index], X[test_index]
        Y_train, Y_test = Y[train_index], Y[test_index]

        Y_pred = [local_linear_regression(x, X_train, Y_train, bandwidth
            ) for x in X_test]
        mse_values.append(mean_squared_error(Y_test, Y_pred))

    return np.mean(mse_values)

# Testing different bandwidths with k-fold cross-validation
bandwidths = [13, 15, 20, 25, 30, 35, 40]
k = 5  # Number of folds

k_fold_mse_values = []

for bandwidth in bandwidths:
    mse = k_fold_cv_llr(lamb_no, bandwidth, k=k)
    k_fold_mse_values.append(mse)

# Plotting
plt.figure(figsize=(10, 6))
plt.plot(bandwidths, k_fold_mse_values, marker=’o’)
plt.title(f’K-Fold CV (k={k}) MSE vs Bandwidth for Lamb 612 (Local
   Linear Regression)’)
plt.xlabel(’Bandwidth’)
plt.ylabel('K-Fold CV MSE')
plt.grid(True)
plt.show()

# Finding the optimal bandwidth
optimal_bandwidth_k_fold = bandwidths[np.argmin(k_fold_mse_values)]
print(f'Optimal Bandwidth based on {k}-Fold CV for Local Linear
    Regression: {optimal_bandwidth_k_fold}')

Result: Optimal Bandwidth based on 5-Fold CV for Local Linear Regression
    : 20
```

Growth Curve of Lamb 612 - Section 4.2.3, Figure 4.6

```
# Define the lamb number and the age range for estimation
lamb_no = 612
X = np.arange(0, 430, 1)  # Estimating weights from day 0 to 429

# Bandwidths to try
bandwidths = [20, 30, 40]

# Plotting the actual data and the estimated curves for different
    bandwidths
plt.figure(figsize=(12, 6))

# Filtering the data for lamb 612
lamb_data = final_df2[final_df2['No.'] == lamb_no]

# Plotting the actual data points
plt.plot(lamb_data['Age (Days)'], lamb_data['Weight'], 'o', label=f'
    Actual Data (Lamb {lamb_no})', color='black')

# Plotting the estimated growth curves for different bandwidths
colors = ['red', 'green', 'blue']
for i, bandwidth in enumerate(bandwidths):
    estimated_weights_llr = estimate_weight_curve_llr(lamb_no, X,
        bandwidth)
    plt.plot(X, estimated_weights_llr, label=f'Estimated Curve (
        Bandwidth = {bandwidth})', color=colors[i], linestyle='--')
```

```
plt.title('Estimated Growth Curves of Lamb 612 with Different Bandwidths
    For LLR')
plt.xlabel('Age (Days)')
plt.ylabel('Weight')
plt.legend()
plt.grid(True)
plt.show()
```

Figure 4.7

```
# Function to plot the growth curve
def plot_growth_curve(df, lamb_no, age_range):
    # Predict weights for the lamb
    predicted_weights = predict_weights_for_lamb(df, lamb_no, age_range)

    # Actual data points for the lamb
    lamb_data = df[df['No.'] == lamb_no]

    # Plotting the actual data points
    plt.scatter(lamb_data['Age (Days)'], lamb_data['Weight'], label=f'
       Lamb {lamb_no} Actual', alpha=0.6)

    # Plotting the predicted growth curve
    plt.plot(age_range, predicted_weights, linestyle='--', color='red',
       label=f'Lamb {lamb_no} Predicted')

    plt.title(f'Growth Curve of Lamb {lamb_no}')
    plt.xlabel('Age (Days)')
    plt.ylabel('Weight')
    plt.legend()
    plt.grid(True)
    plt.show()

#Plot for Lamb 612
lamb_no = 612
age_range = np.arange(90, 450, 2)
```

Figure 4.8

```
# Function to predict weights for given litter size, gender, and
  pedigree type
def predict_general_trend(age_range, litter_size, gender, type_pedigree,
    model, poly):
    # Prepare the input data for prediction
    X_pred = np.array([[age, litter_size, gender, type_pedigree] for age
        in age_range])
    X_poly_pred = poly.transform(X_pred)
    predicted_weights = model.predict(X_poly_pred)

    return predicted_weights

# Function to estimate weights for a given lamb and plot the actual and
  predicted growth curve
def plot_individual_lambs(lamb_ids, age_range, model, poly):
    plt.figure(figsize=(14, 8))

    for lamb_no in lamb_ids:
        # Extract actual data for the lamb
        lamb_data = comb_mul_df_o[comb_mul_df_o['No.'] == lamb_no].
            dropna(subset=['Weight'])

        # Plot actual data points
        plt.plot(lamb_data['Age (Days)'], lamb_data['Weight'], 'o',
            label=f'Actual Data (Lamb {lamb_no})')

    # Since all lambs are female, have a 2 litter size, and are pedigree
        type, we can use one prediction curve
    litter_size = 2
    gender = 0  # Female
    type_pedigree = 1  # Pedigree

    # Predictting the growth curve using the defined model and
       polynomial features
    predicted_weights = predict_general_trend(age_range, litter_size,
       gender, type_pedigree, model, poly)

    # Plotting the predicted growth curve (same for all these lambs)
    plt.plot(age_range, predicted_weights, linestyle='--', color='red',
        label='Predicted Growth Curve')

    plt.title('Actual Data and Predicted Growth Curve for Lambs')
    plt.xlabel('Age (Days)')
    plt.ylabel('Weight')
    plt.legend()
    plt.grid(True)
    plt.show()

# Define the lamb IDs to plot
lamb_ids = [711, 825, 829, 931]

# Define the age range for prediction
age_range = np.arange(90, 425, 2)  # Age range from 90 to 400 days

# Plot the growth curves and actual data for the specified lambs
plot_individual_lambs(lamb_ids, age_range, model, poly)
```

Section-4.3.3, Figure 4.9,4.10

```
# Function to plot general growth trends by type_pedigree
def plot_general_trends(model, poly, age_range):
    # Define litter sizes, genders, and pedigree types for the general
        trends
    litter_sizes = [1, 2, 3,4]
    genders = [0]  # Assuming 0 for female, 1 for male
    pedigree_types = [0, 1]  # Assuming 0 for Commercial, 1 for Pedigree

    fig, axes = plt.subplots(1, 2, figsize=(20, 8), sharey=True)
    fig.suptitle('General Growth Trends for Lambs by Pedigree Type',
        fontsize=16)

    for i, type_pedigree in enumerate(pedigree_types):
        for litter_size in litter_sizes:
            for gender in genders:
                # Predict weights for the general trend
                predicted_weights = predict_general_trend(age_range,
                    litter_size, gender, type_pedigree, model, poly)
                # Plot the predicted growth curve on the corresponding
                   subplot
                gender_label = 'Male' if gender == 1 else 'Female'
                pedigree_label = 'Pedigree' if type_pedigree == 1 else '
                   Commercial'
                label = f'Litter Size {litter_size}, {gender_label}'
                axes[i].plot(age_range, predicted_weights, linestyle
                   ='--', label=label)

        # Labeling each subplot
        axes[i].set_title(f'{pedigree_label} Lambs')
        axes[i].set_xlabel('Age (Days)')
        axes[i].set_ylabel('Weight')
        axes[i].legend()
        axes[i].grid(True)

    plt.show()

age_range = np.arange(90, 425, 2)  # Age range from 90 to 400 days
plot_general_trends(model, poly, age_range)
```

Bibliography

[1] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

[2] Wright, N. (2021). *Key Performance Indicators of Ewe Productivity: Importance of Ewe Body Condition Score and Liveweight on Pregnancy Outcomes and Lamb Performance to Weaning* (Doctoral dissertation, University of Nottingham).

[3] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. `https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf`, accessed 18 August 2024.

[4] Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test*, 18(1), 1–43.

[5] Brown, C. H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, 143–155.

[6] Härdle, W. (1990). *Applied nonparametric regression* (No. 19). Cambridge University Press.

[7] Dakhil, M. A., & Hussain, J. N. (2021, March). A Comparative Study of Nonparametric Kernel Estimators with Gaussian Weight Function. In *Journal of Physics: Conference Series* (Vol. 1818, No. 1, p. 012058). IOP Publishing.

[8] Demir, S., & Toktamiş, Ö. (2010). On the adaptive Nadaraya-Watson kernel regression estimators. *Hacettepe Journal of Mathematics and Statistics*, 39(3), 429–437.

[9] Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression* (No. 12). Cambridge University Press.

[10] Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596–610.

[11] Thorson, J. T., & Taylor, I. G. (2014). A comparison of parametric, semi-parametric, and non-parametric approaches to selectivity in age-structured assessment models. *Fisheries Research*, 158, 74–83.

[12] Zimmerman, D. L., Núñez-Antón, V., Gregoire, T. G., Schabenberger, O., Hart, J. D., Kenward, M. G., ... & Vieu, P. (2001). Parametric modelling of growth curve data: An overview. *Test*, 10(1), 1–73.

[13] Verbeke, G., Molenberghs, G., & Verbeke, G. (1997). *Linear Mixed Models for Longitudinal Data* (pp. 63-153). Springer New York.

[14] Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons.

[15] Müller, H. G., & Stadtmüller, U. (2005). *Generalized functional linear models*.

[16] Fan, J. (2018). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge.

[17] Zhang, Y. (2014, July). Bandwidth selection for Nadaraya-Watson kernel estimator using cross-validation based on different penalty functions. In *International Conference on Machine Learning and Cybernetics* (pp. 88-96). Berlin, Heidelberg: Springer Berlin Heidelberg.