Danmarks Tekniske Universitet

DTU

Analysis of Correlated Data
Assignment 5

Marios-Dimitrios Lianos

s233558

December 11, 2023

Summary

For the requirements of Assignment 5 in the Analysis of Correlated Data: Mixed Linear Models class, a detailed statistical analysis was conducted. The primary objective of this study was to investigate the impact of different diets and the progression of time on the protein percentage in cow milk. The study aimed to discern how varying diets influence milk protein content over a period and to quantify these effects using appropriate statistical models.

The analytical approach of this study was divided into three distinct phases:

Model Diagnostics and Selection: The initial phase focused on examining various mixed-effects models to accurately represent the data.

Statistical Analysis: The central part of the analysis involved applying the chosen mixed-effects model to understand the fixed effects of diet and time on milk protein percentage. This phase included evaluating the significance of these effects, conducting post-hoc pairwise comparisons, and estimating marginal means to interpret the fixed effects in a practical context.

Interpretation and Conclusions: The final phase synthesized the analytical findings into meaningful insights.

Contents

1 Introduction

This study is dedicated to examining the influence of different diets and the progression of time on the protein percentage in cow milk. The research involved monitoring and recording the protein content in milk from cows subjected to three distinct diets over a period of several weeks. The aim is to develop a comprehensive understanding of how dietary changes and time factors affect milk composition, with a particular focus on protein percentage. This knowledge is crucial for optimizing dairy farming practices, improving milk quality, and enhancing overall dairy production efficiency.

2 Exploratory Data Analysis

2.1 Variables in the Dataset

- `Cow`: Categorical variable representing each cow in the study. This variable is crucial for accounting for individual differences among cows, which might affect milk protein percentages.
- `Diet`: Categorical variable that denotes the type of diet fed to each cow. The inclusion of diet in the analysis allows for the examination of how different feeding regimes influence the protein percentage in cow milk.
- `Time`: Continuous variable representing the time in weeks over which the protein percentage measurements were taken. This variable captures the temporal aspect of the study and is essential for analyzing how the protein percentage in milk changes over time.
- `Protein`: This is a continuous variable representing the measured protein percentage in cow milk. It's the primary response variable in the study, used to assess the impact of diet and time on milk quality.

In our analysis, the factor `Cow`, representing individual cows, is treated as a random effect due to the inherent variability among different cows and the interest in generalizing our findings beyond the observed sample. The factor `Diet`, denoting the type of diet, is a fixed effect as we are specifically interested in the effect of different diets on milk protein percentage. `Time`, indicating the week of measurement, is also treated as a fixed effect to understand the temporal dynamics of protein percentage.

In terms of balance, the dataset is balanced meaning each unique combination of these three factors has an equal count. As for the structure, `Cow` is nested within `Diet`, as each cow is subjected to one type of diet, while `Time` is crossed with both `Cow` and `Diet`, considering that measurements are taken over time for each cow on each diet.

2.2 Exploratory Diagrams

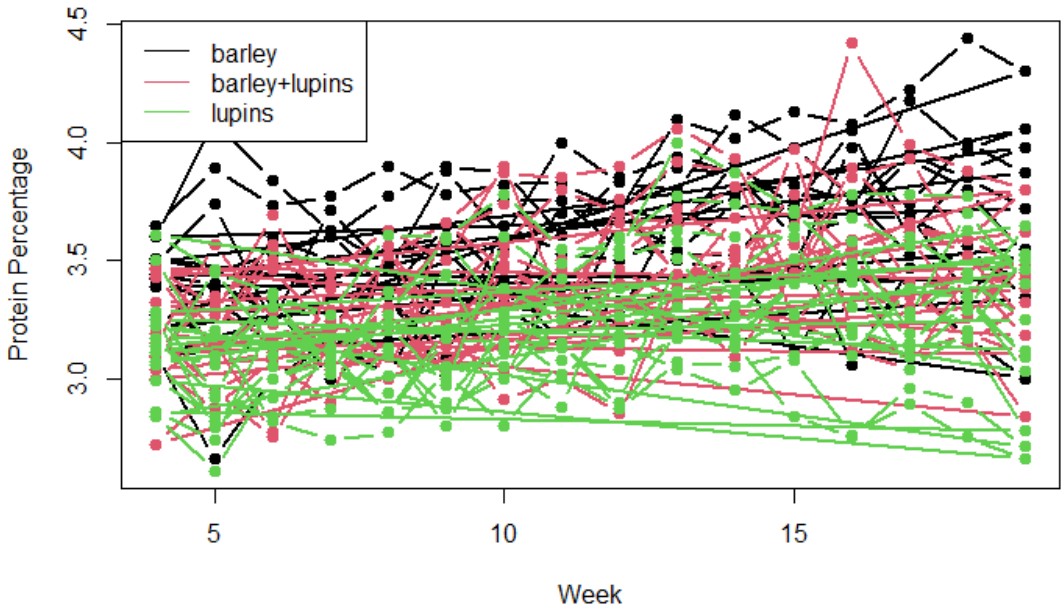

Figure 1: Protein Percentage Over Time for Each Diet

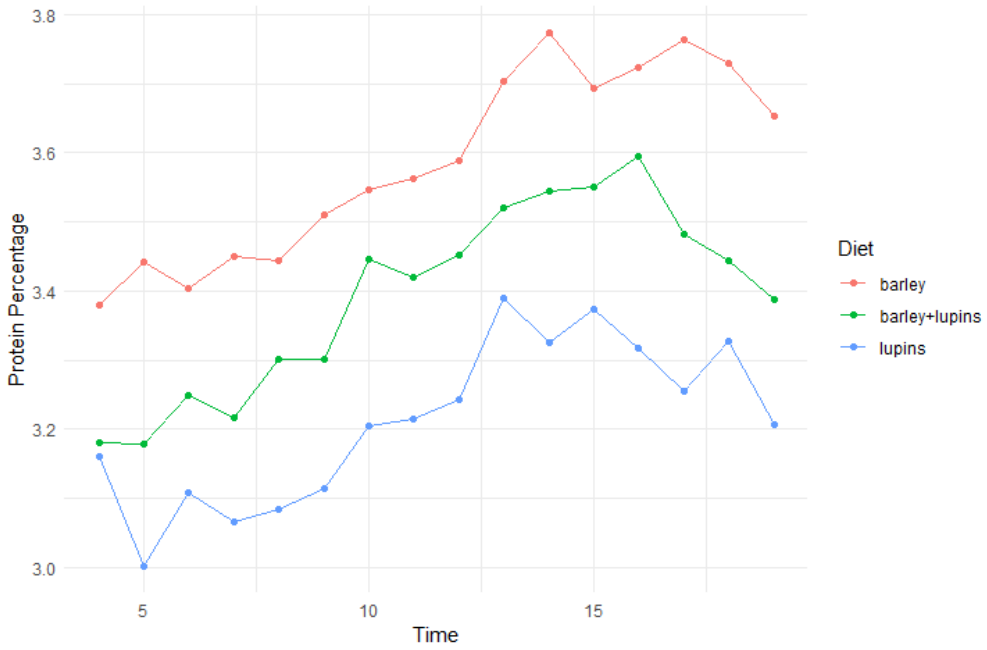

Figure 2: Mean Protein Level by Diet and Time

In Fig: 2 there is a notable difference in the protein percentage trends associated with each diet, which might suggest that barley contributes to a higher protein percentage in the milk compared to lupins.

Fig: 3 suggests that the distribution of protein percentages in the dataset is roughly bell-shaped but shows some signs of being skewed to the right. The peak of the distribution is around 3.5%.

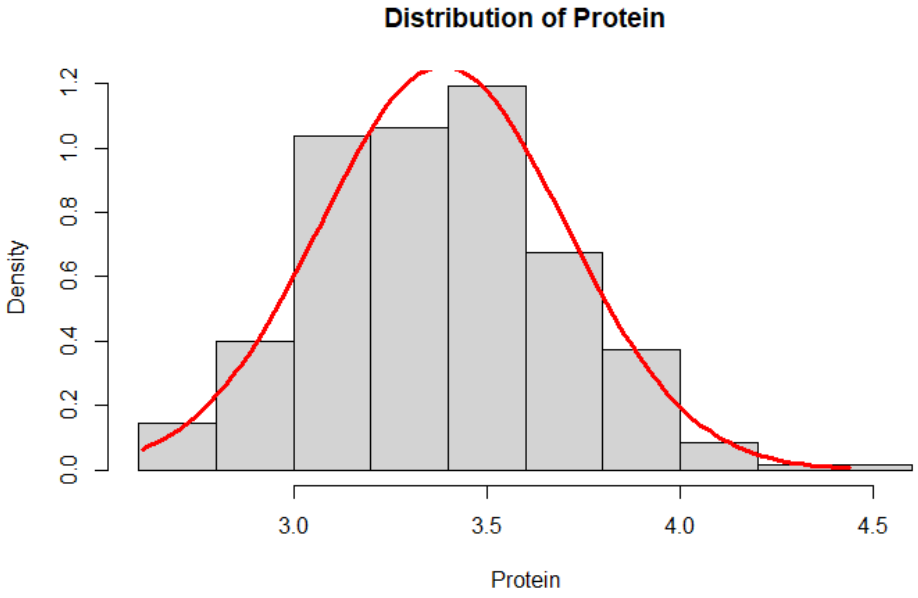

Figure 3: Histogram for Protein

2.3 Exploratory Model Analysis

One of the primary challenges in our analysis stems from the potential violation of the independence assumption typically associated with standard statistical methods. In datasets like ours, where multiple measurements are taken from the same individual, it is reasonable to expect a degree of positive correlation between these measurements. This correlation implies that observations from the same individual are likely to be more similar to each other than to observations from different individuals. Additionally, the degree of correlation between measurements may vary with time; measurements taken closer together in time are often more strongly correlated than those taken further apart. These characteristics necessitate a careful approach to the analysis that accounts for such individual correlations.

To address these intricacies, we have chosen to implement the Gaussian model of spatial correlation in our analysis. This model is particularly suited for data where measurements are temporally correlated, as it accommodates the varying degrees of correlation between

observations based on their temporal proximity. The Gaussian spatial correlation model assumes that measurements taken closer together in time will exhibit a higher degree of correlation, diminishing as the time gap widens. This approach not only helps in modeling the inherent correlation structure within our data effectively but also enhances the robustness and reliability of our statistical inferences, ensuring they are well-suited to the nuanced nature of our dataset.

The mathematical representation of the initial complex model:

$$Y_{ij} = \beta_0 + \beta_1 \times \text{Diet}_{ij} + \beta_2 \times \text{Time}_{ij} + \beta_3 \times \text{Diet}_{ij} \times \text{Time}_{ij} + b_{0i} + b_{1i} \times \text{Diet}_{ij} + b_{2i} \times \text{Time}_{ij} + \epsilon_{ij} \quad (a)$$

Y_{ij} is the protein percentage for the i-th cow at the j-th time point, β_0 the overall intercept. β_1, β_2, and β_3 are the fixed effect coefficients for Diet, Time, and their interaction. Diet_{ij} and Time_{ij} are the values of the Diet and Time variables for the i-th cow at the j-th time point. b_{0i} is the random intercept for the i-th cow. b_{1i} and b_{2i} are the random slopes for Diet and Time for the i-th cow respectively. ϵ_{ij} is the residual error.

2.4 Model Reduction

The next procedure is a way to examine which terms can be dropped off from (a). When considering which terms to exclude from a mixed-effects model, we typically start by examining the most complex terms and progressively move to simpler terms, ensuring at each step that the model still adequately fits the data.

The first term that we will examine is **(Diet + Time | Cow)** which gives **'log Lik.'** 0.5 (df=18). As a consequence, we can remove the random slopes of *Diet* and *Time* within *Cow*.

Next we will compare our model, with a model without the **Gaussian spatial correlation**.

Model	logLik	AIC	BIC	Pr(>Chisq)
model2	42.56931	-67.13861	-27.77873	
model3	5.79511	4.40978	39.39634	<.0001
'log Lik.' 0				

Table 1: Anova for model2, model3

In Tab: 1 the model with the Gaussian spatial correlation (model2) has a significantly lower AIC and higher log-likelihood than the model without the spatial correlation (model3). The likelihood ratio test (L.Ratio) is 73.54839 with a very low p-value (<.0001), which suggests that the spatial correlation structure is a significant component of the model and should not be removed.

Finally, we will examine the significance of the interaction term of the fixed effects $Diet : Time$. If the test shows that we can reduce the model, the next step would be to validate the covariance structure. Since comparing models with different fixed effects using REML (Restricted Maximum Likelihood) is not appropriate we will use ML (Maximum Likelihood) instead of REML for fitting the models.

Model	logLik	AIC	BIC	Pr(>Chisq)
model2	62.83990	-107.6798	-68.22824	
model4	49.91607	-109.1630	-78.47848	0.2841

Table 2: Anova for model2, model4

In Table: 2 the interaction between Diet and Time is not statistically significant in explaining the variability in protein percentage, and therefore, the simpler model without this interaction term can be used.

Covariance structure

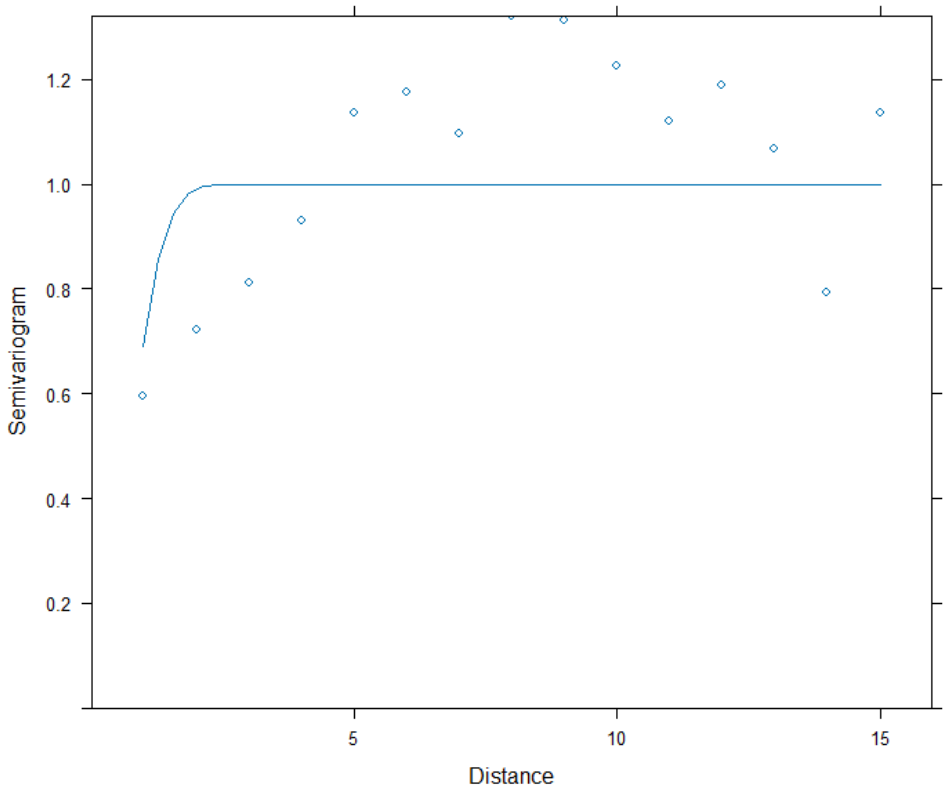

Figure 4: Semivariogram for the Model with the $Diet : Time$ Interaction

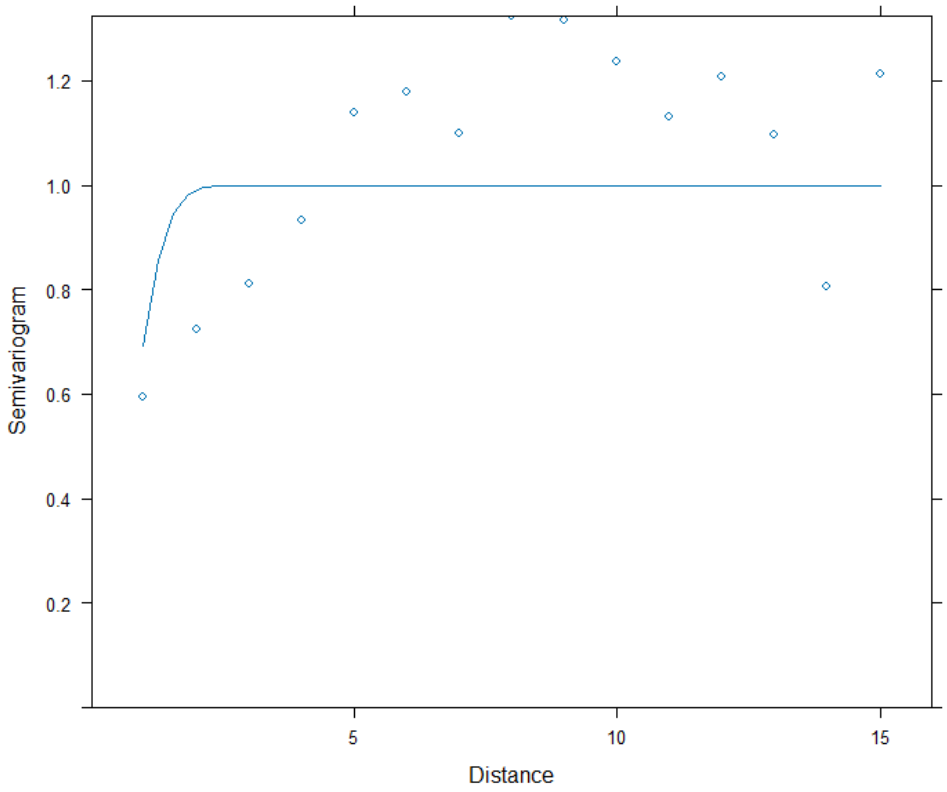

Figure 5: Semivariogram for the Model without the $Diet : Time$ Interaction

Given that both semivariograms Fig: 4 and Fig: 5 appear to be identical and display a similar pattern, the removal of the $Diet : Time$ interaction doesn't seem to have altered the covariance structure captured by the Semivariogram.

Final Model

The final examination will be if the reduced model fits the data better if we use a model with Exponential correlation than the Gaussian spatial correlation model.

Model	logLik	AIC	BIC	L.Ratio	Pr(>Chisq)
$model_{Gauss}$	49.91607	-85.83214	-55.19505		
$model_{Exp}$	74.75782	-133.51563	-98.50181	49.68349	<.0001

Table 3

From Table: 3 the likelihood ratio test (L.Ratio) comparing the two models gives a value of 49.68349, with a p-value that is less than 0.0001. This very low p-value strongly suggests that $model_{Exp}$ with the exponential correlation structure and nugget effect provides a significantly better fit to the data than $model_{Gauss}$ with the Gaussian spatial correlation.

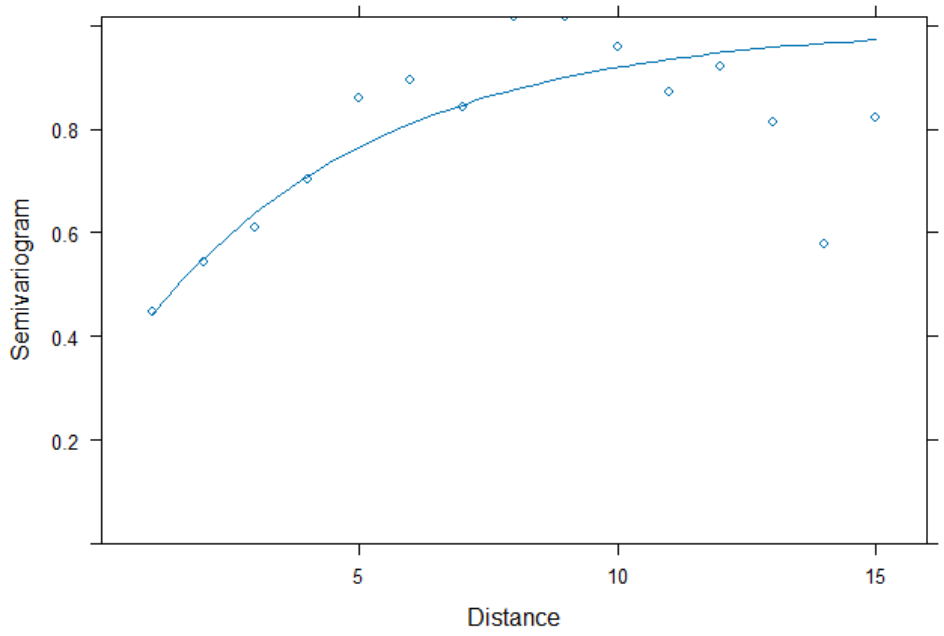

Figure 6: Semivariogram for the Exponential Model

Model	logLik	AIC	BIC	L.Ratio	Pr(>Chisq)
$model_{nugget}$	74.75782	-85.83214	-98.50181		
$model$	64.71681	-115.43363	-84.79654	20.082	<.0001

Table 4: Comparison of the model with a nugget effect and the model without

In Table: 4 the statistical evidence points to the conclusion that $model_{nugget}$, which accounts for extra variability at zero distance (or the nugget effect), is a more appropriate model for the data. The nugget effect here is capturing additional variability that is not explained by the exponential correlation structure alone, which could represent measurement error or other sources of variability at the smallest scale.

Final Representation of the Model

The linear mixed-effects model for the protein percentage (*protein*) with an exponential correlation structure:

$$protein_{ij} = \beta_0 + \beta_1 \times \text{Diet}_{ij} + \beta_2 \times \text{Time}_{ij} + b_{0i} + \epsilon_{ij} \tag{1}$$

where:

- $protein_{ij}$ is the protein percentage for the i-th cow at the j-th time point.
- β_0 is the overall intercept.
- β_1, and β_2, are the fixed effect coefficients for Diet and Time, respectively.
- Diet_{ij} and Time_{ij} are the values of the Diet and Time variables for the i-th cow at the j-th time point.
- b_{0i} is the random intercept for the i-th cow, assumed to be normally distributed with mean 0 and variance $\sigma^2_{b_0}$.
- ϵ_{ij} is the residual error for the i-th cow at the j-th time point, which follows an exponential spatial correlation structure with a nugget effect.

The exponential spatial correlation structure for the residual errors, including a nugget effect, is given by:

$$\text{corr}(\epsilon_{ij}, \epsilon_{ik}) = \tau + (1 - \tau) \exp\left(-\frac{|\text{Time}_{ij} - \text{Time}_{ik}|}{\phi}\right) \tag{2}$$

where:

- τ represents the nugget effect, capturing unexplained variability at zero distance.
- ϕ is a parameter governing the range of the spatial correlation.

3 Statistical Analysis

3.1 Model Diagnostics

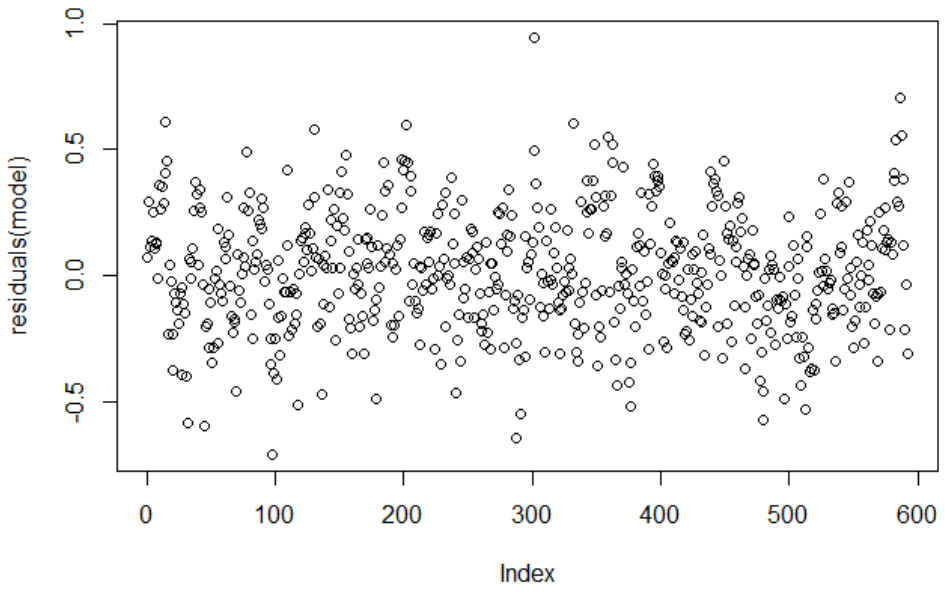

Figure 7: Residuals vs Index Plot

In Fig: 7 the residuals appear to be randomly scattered around the zero line without any discernible pattern, which suggests that the variances are constant and that there are no obvious trends or periodicity. This is a good indication that the model is capturing the data's structure well, although a few outliers are present.

In Fig: 8 the points in the Q-Q plot closely follow the reference line, particularly in the center of the distribution, which indicates that the residuals are approximately normally distributed.

In Fig: 9 the residuals are randomly distributed around the zero line with no clear pattern as the fitted values increase, which supports the assumption of homogeneity of variance. The absence of a clear pattern also suggests that the model does not suffer from non-linearity or omitted variable bias.

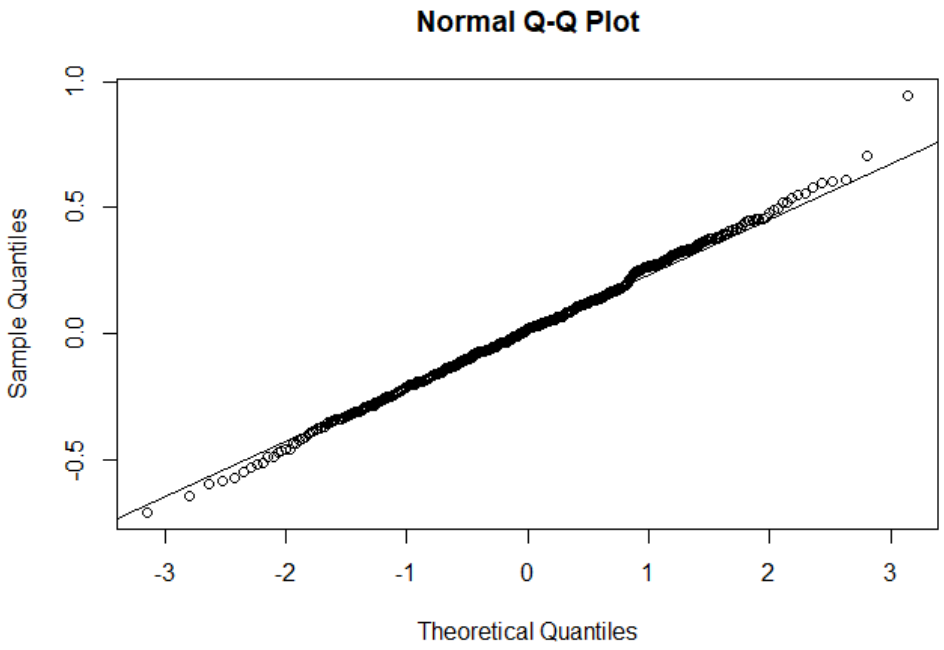

Figure 8: Normal Q-Q Plot

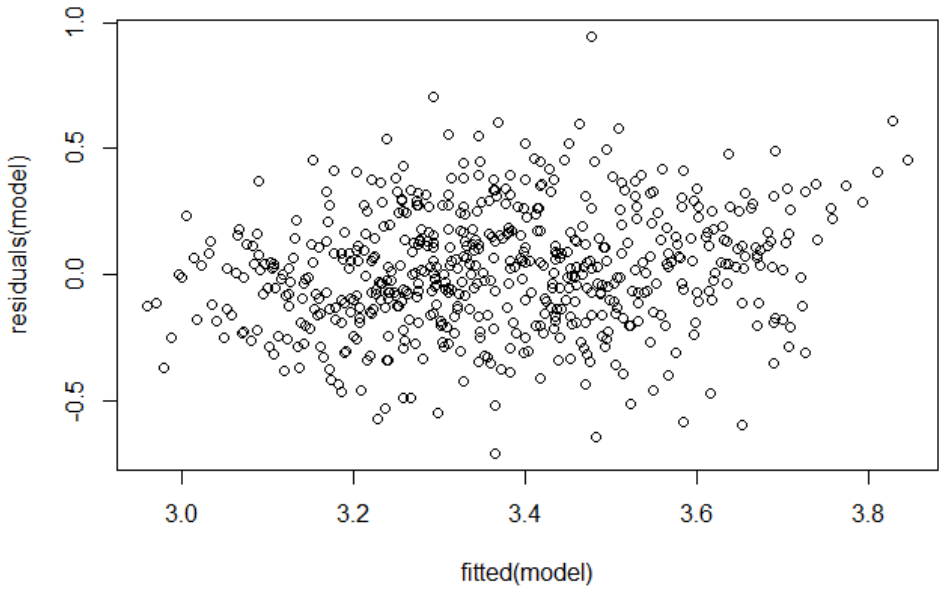

Figure 9: Residuals vs Fitted Values Plot

3.2 Results

The approximate confidence intervals for the parameters estimated in the model, can be very useful for understanding the precision of the estimates.

Level : *Cow*	Lower	Estimate	Upper
sd((Intercept))	0.04014545	0.09885738	0.2434343

Table 5: The standard deviation of the random intercept for Cow

In Table: 5 There is some variability in the baseline protein percentage across cows, but the exact amount of this variability is uncertain. The relatively wide confidence interval suggests that while you can be confident that there is some variation between cows, the precise degree of this variability is less certain.

	Lower	Estimate	Upper
Range	2.3890735	4.6470322	9.0390304
Nugget	0.2130171	0.3096067	0.4262706

Table 6: Correlation Structure

In Table: 6 the confidence interval for the range is wide, indicating that while the model estimates that measurements are correlated over some time range, the exact extent of this range is uncertain. The nugget effect is significant and represents variability at the smallest scale (such as measurement error) that is not explained by the model. The confidence interval for the nugget effect suggests that there is some measurement error or micro-scale variability, but again, the precise amount is uncertain.

Within-group standard error		
Lower	Estimate	Upper
0.2176601	0.2518048	0.2913058

Table 7: Within-Group Standard Error

In Table: 7 the residuals have a certain amount of variability around the predicted values, indicating the presence of noise in the measurements. The confidence interval provides an estimate of the range in which the true standard error of the residuals is likely to fall.

Term	Value	Std.Error	DF	p-value
(Intercept)	3.367698	0.06454818	554	<0.0001
barley+lupins	-0.206962	0.06891549	34	0.005
lupins	-0.369585	0.06891549	34	<0.0001
Time	0.017795	0.00347163	554	<0.0001

Table 8: Anova Table for Fixed Effects

Overall, Table: 8 provides evidence that both *Diet* and *Time* are significant predictors of protein percentage in cow milk, with different diets having distinct impacts compared to a baseline diet, and a general trend of increasing protein percentage over time. The statistical significance of these predictors is confirmed by their p-values.

3.3 Post-hoc Analysis

In these tables the estimated marginal means and pairwise comparisons are specifically for the 11.5-week time point. This time point has been selected to examine the effect of diet on protein percentage at a specific moment in the study period. It provides insight into how the diets compare at a particular stage rather than over the entire study duration.

This focused analysis helps understand the diet effect at a mid-point, which can be particularly relevant if we're interested in how dietary impacts evolve over time or if there's a specific time when dietary effects are most pronounced.

Diet	Estimate	SE	Df	Lower	Upper
barley	3.57	0.0507	36	3.47	3.68
barley+lupins	3.37	0.0467	34	3.27	3.46
lupins	3.20	0.0467	34	3.11	3.30

Table 9: Post-hoc

Table: 9, shows that cows on a barley diet have the highest estimated protein percentage in their milk, followed by those on barley+lupins, and finally those on lupins.

contrast	estimate	SE	df	t.ratio	p.value
barley - (barley+lupins)	0.207	0.0689	34	3.003	0.0135
barley - lupins	0.370	0.0689	34	5.363	<.0001
(barley+lupins) - lupins	0.163	0.0660	34	2.465	0.0484

Table 10: Pairwise Comparison of different *Diets*

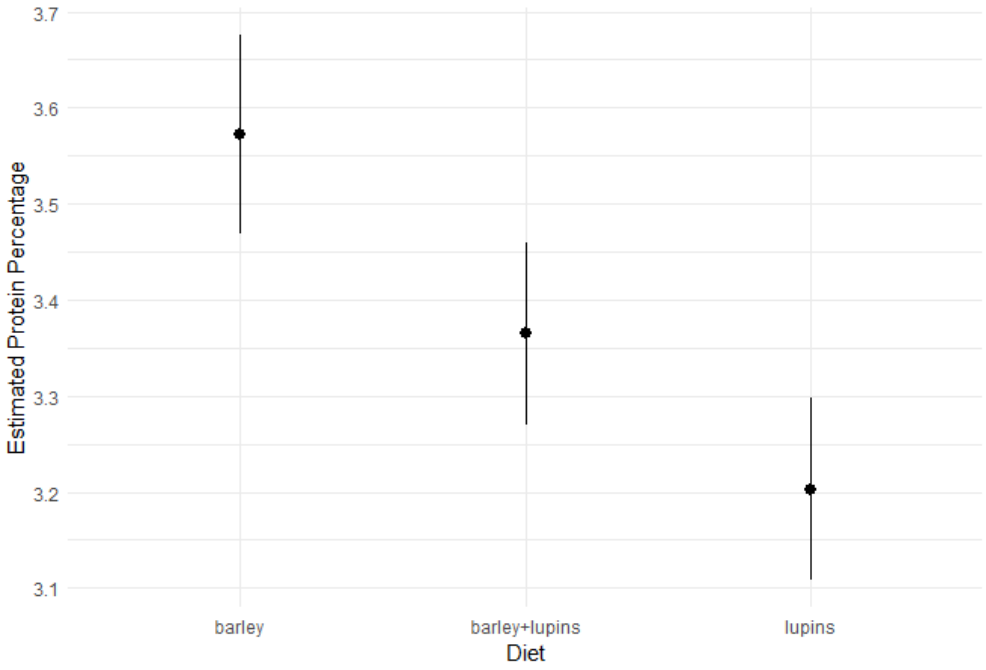

Figure 10: Confidence intervals for the estimated marginal means, T = 11.5

3.4 Cross-Validation of the Results

Given the nature and complexity of this type of model, it is recommended that the main conclusions of a given study should be $cross - validated$ with one of the simple methods (Separate analyses for each $time - point$, Analysis of summary statistics, Random effects approach).

$$protein_{ij} = \beta_0 + \beta_1 \times \text{Diet}_{ij} + \beta_2 \times \text{Time}_{ij} + b_{0i} + \epsilon_{ij} \tag{3}$$

$protein_{ij}$ is the protein percentage for the i-th cow at the j-th time point. β_0 is the overall intercept. β_1, and β_2, are the fixed effect coefficients for Diet and Time, respectively. Diet_{ij} and Time_{ij} are the values of the Diet and Time variables for the i-th cow at the j-th time point. b_{0i} is the random intercept for the i-th cow, assumed to be normally distributed with mean 0 and variance $\sigma^2_{b_0}$. ϵ_{ij} is the random error term.

The main difference of models (1)-(3) is that in this random effects approach all measurements on the same individual are assumed equally correlated.

	Estimate	Std. Error	t value
(Intercept)	3.328693	0.053551	62.159
barley+lupins	-0.193387	0.066163	-2.923
lupins	-0.373051	0.066163	-5.638
Time	0.022327	0.001938	11.522

Table 11: Anova Table - Model (3)

The cross-validation process demonstrates that the conclusions regarding the fixed effects (impact of Diet and Time) are robust across different model specifications. This adds confidence to the findings that both the type of diet and the progression of time are important determinants of protein percentage in cow milk.

Both models consistently show that different diets significantly affect protein percentage, with $barley + lupins$ and $lupins$ diets leading to lower protein percentages compared to the baseline diet $barley$. The increase in protein percentage over time is a consistent finding across both models, though the rate of increase varies slightly. The slight differences in the estimates and standard errors between the two models are due to how each model accounts for the correlations and random effects.

4 Conclusions

Diet Effects:

Diet barley: The value (3.367698) is the estimated mean protein percentage when Diet and Time are at their reference levels. The p-value (<0.0001) indicates that the diet is significantly different from zero.

Diet barley+lupins: The negative estimate (-0.206962) suggests that this diet results in a lower protein percentage compared to the baseline diet (barley), and the effect is statistically significant.

Diet lupins: A more substantial negative effect (-0.369585) implies a larger decrease in protein percentage with the lupins diet compared to the baseline, also statistically significant.

Time Effect: A positive estimate (0.017795) with a high t-value indicates that protein percentage increases over time.

Conclusion: Cows on a barley diet have the highest estimated protein percentage in their milk, followed by those on barley+lupins and those on lupins.

5 Appendix A

```
# Libraries
library(nlme) #!!!!****
library(ggplot2)
library(emmeans)
library(plyr)
library(dplyr)
library(MASS)
library(car)
library(multcomp)
library(lmerTest)
library(Matrix)
library(lme4)

# Reading in data:
# In a study on the protein percentage in milk, cows were given three types of diets. 37 cows
# were randomly assigned to a diet, and measurements of the protein percentage were taken
# for each cow for 19 consecutive weeks (the first 3 weeks were a settling-in period and are not
# included in the data). The aim of the study was to investigate how the three diets affect the
# milk production over time in terms of the protein percentage.
# The variable [Cow] identifies the cow, [Diet] denotes the assigned diet, [Time] denotes the week
# (4-19) the measurement was taken, and [Protein] denotes the measured protein percentage.
# The data is available in the file assignment5.txt.

# we will treat Time as continous variable

assignment5 <- read.table("assignment5.txt", header = TRUE, sep = "\t")
assignment5$Cow <- as.factor(assignment5$Cow)
assignment5$Diet <- as.factor(assignment5$Diet)
assignment5$Time <- as.numeric(assignment5$Time)
assignment5$Timef <- NULL
summary(assignment5)

# Check for missing values
any(is.na(assignment5))

#Check if Factors are Balanced
assignment5 %>%
```

```
  group_by(Cow, Diet,Time) %>%  # Use actual factor names here
  summarise(count = n(), .groups = "drop") %>%
  distinct(count) %>%
  nrow() -> num_unique_counts
if (num_unique_counts == 1) {
  print("The dataset is balanced.")
} else {
  print("The dataset is unbalanced.")
}

#---------------Diagrams--------------#
ggplot(assignment5, aes(x=Time, y=protein, group=Cow, colour=Diet)) +
    geom_line()


# Calculate the mean protein level for each combination of Diet and
   Time
mns <- assignment5 %>%
  group_by(Diet, Time) %>%
  summarise(protein = mean(protein, na.rm = TRUE), .groups = "drop")
ggplot(mns, aes(x = Time, y = protein, group = Diet, colour = Diet))
   +
  geom_point() +
  geom_line() +
  labs(x = "Time", y = "Protein Percentage") +
  theme_minimal() +
  scale_colour_discrete(name = "Diet")



# Plotting the protein percentage over time for each cow under
   different diets.
with(assignment5, plot(Time, protein, xlab='Week', ylab='Protein
   Percentage', pch=20))
unique_diets <- unique(assignment5$Diet)
for(d in unique_diets){
    temp <- assignment5[assignment5$Diet == d,]
    with(temp, lines(Time, protein, type="b", col=which(unique_diets
       == d), lwd=2, pch=19))
}
legend("topleft", legend=unique_diets, lty=1, col=1:length(unique_
   diets))


# Histogram for Protein
f <- function(x) {
```

```
  dnorm(x, mean = mean(assignment5$protein), sd = sd(assignment5$
    protein))
}
hist(assignment5$protein, xlab='Protein', probability=T, main = "
  Distribution of Protein")
curve(f, from = min(assignment5$protein), to = max(assignment5$
  protein), lwd=3, col="red", add=T)



#----------------------Gaussian model of spatial correlation
  ------------------------------------#


model <- lme(fixed = protein ~ Diet * Time,  # Interaction between
  fixed effects
             random = ~ Diet + Time | Cow,  # Random slopes for fixed
                 effects within 'Cow'
             correlation = corGaus(form = ~ Time | Cow),
             data = assignment5)
summary(model)

#----------------------Model Reduction-------------------------#

model2 <- update(model, .~. - (Diet + Time | Cow))
anova(model, model2)
0.5*(1-pchisq(2*(logLik(model)-logLik(model2)),1))

model <- lme(fixed = protein ~ Diet * Time,
                        random = ~ 1 | Cow,
                        correlation = corGaus(form = ~ Time | Cow),
                        data = assignment5)
summary(model)

model3 <- update(model, correlation = NULL)
anova(model, model3)
0.5*(1-pchisq(2*(logLik(model)-logLik(model3)),1))


model4 <- update(model, .~. - (Diet:Time))
model_ml <- update(model, method = "ML")
model4_ml <- update(model4, method = "ML")
anova(model_ml, model4_ml)


plot(Variogram(model, form= ~ Time | Diet, data=assignment5))
```

```
plot(Variogram(model4, form= ~ Time | Diet, data=assignment5))


# Final model

#Gaussian spatial correlation model
model <- lme(fixed = protein ~ Diet + Time,
                        random = ~ 1 | Cow,
                        correlation = corGaus(form = ~ Time | Cow),
                        data = assignment5)

plot(Variogram(model, form= ~ Time | Diet, data=assignment5))

#Exponential correlation model with the nugget effect
model2 <- lme(protein ~ Diet + Time,
                     random= ~ 1 | Cow,
correlation=corExp(form= ~ Time | Cow, nugget = TRUE),
data = assignment5)

anova(model,model2)

plot(Variogram(model2, form = ~as.numeric(Time) | Cow, data =
   assignment5))

model3 <- lme(protein ~ Diet + Time,
                     random= ~ 1| Cow,
correlation=corExp(form= ~Time | Cow),
data=assignment5)

anova(model2,model3)

#-------------------Model Diagnostics-----------------------#

#Exponential correlation model with the nugget effect
model <- lme(protein ~ Diet + Time,
                     random= ~ 1 | Cow,
correlation=corExp(form= ~ Time | Cow, nugget = TRUE),
data = assignment5)
plot(Variogram(model, form = ~as.numeric(Time) | Cow, data =
   assignment5))

# Checking residuals of the mixed-effects model
plot(residuals(model))

# Checking normality of residuals
qqnorm(residuals(model))
```

```
qqline(residuals(model))

# Plotting residuals vs. fitted values
plot(fitted(model), residuals(model))

plot(augPred(model), addData = TRUE)

# Checking for influential observations
influence_measures <- influence(model)
plot(influence_measures, id = 0.05)  # Plot with 5% cutoff for influential points

#----------------------------Statistical Analysis----------------------------#

model <- lme(protein ~ Diet + Time,
                     random= ~ 1 | Cow,
correlation=corExp(form= ~ Time | Cow, nugget = TRUE),
data = assignment5)
summary(model)
anova(model)

intervals(model, which = "var-cov")

#----------------------------Post-hoc----------------------------------#

#LS-means:
emm <- emmeans(model, "Diet", by="Time", data=assignment5)
print(summary(emm))

# Pairwise
pairwise_emm <- pairs(emm, adjust = "tukey")
print(pairwise_emm)

emms <- emmeans(model, specs = ~ Diet, at = list(Time = 11.5))
emms_df <- as.data.frame(emms)
emm_plot <- ggplot(emms_df, aes(x = Diet, y = emmean, ymin = lower.CL, ymax = upper.CL)) +
            geom_pointrange() +
            xlab("Diet") +
            ylab("Estimated Protein Percentage") +
            ggtitle("Confidence Intervals for Protein Percentage by Diet at Time = 11.5 weeks") +
            theme_minimal()

```

```
202 print(emm_plot)
203
204
205 #----------------Cross-Validation----------------#
206
207 model <- lmer(protein ~ Diet + Time + (1|Cow), data = assignment5)
208 summary(model)
209 anova(model)
```