Data Analysis 3

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

November 24, 2024

Introduction

In this paper we will be looking at data related to calves. The data comes from an experiment designed to study the impact dietary treatments given to pregnant heifers had on the development of the calves. The study was conducted over a three year period and involved three different dietary treatments given to select groups of heifers in the final trimester. In total the data has 22 variables for 120 entires, though some data points are missing.

For more information on the experiment, the data, or any other files used in this paper see our Github page which can be found at https://github.com/RyanLalicker/Data-Analysis-2-STAT-325-825. The coding languages used in the paper are R and SAS. The corresponding code can be found in *Appendix A - R Code* and *Appendix B - SAS Code* respectively.

Variables

As mentioned above the experiment used three different dietary treatments. These were DDG, CON, and MET. For the first two trimesters the heifers were given one of seven developmental treatments, found in `Development.Treatment`, and then in the final trimester the each was given one of the three treatments mentioned above. This is recorded in the `Calan.Treatment` column of the dataset.

The heifers were placed into one of four pens by weight, which can be seen in the column `Pen #`. They were then artificially inseminated from an assigned sire, which we will assume was done randomly since the client says weight was not a factor. The sire is represented by the column of the same name and has six unique entries.

Upon the birth of the calves, several measurements were taken. These include the sex of the calf, weights taken at both birth and slaughter, and scores of both the calf's vigor and the ease of birth. The vigor score is on a scale of one to eleven where a score of one is very good and a score of ten or eleven indicates poor vitality for the calf. (Probo (2022)). The ease score goes from one to five where one indicates a quick and easy birth, two means a longer birth, three

means requires some assistance, and four or five indicates more assistance was needed. (Heins (2023)). Note, the variable names in the dataset line up with the descriptions above.

Other variables, such as the id of the calf, length of gestation for the heifer, and postmortem scoring such as hot carcass weight (HCW) are included as well. (Saner (2024)). Note two birthdays are included in the data, `Birth.date` and `Birth.date.1`. These variables will not be used in the models below so no further investigation was done on our part to determine the differences.

The client's main focus is the effect the third trimester treatment and the sex of a calf have on the calf's vigor score, ease of birth score, and final body weight. Therefore, these are the variables we will place more of an emphasis on, while exploring the effect some of the other variables may have.

Missing Values

The data contains some missing values. In total 53 rows in the dataset are missing at lease one variable. Figure 1 shows which columns have the most missing data. As we can see the values for the variable `DMI`, which according to the USDA represents the dry matter intake for a cow, is missing for two-thirds of the entries. (USDA). Given the number of missing values is this large, it is probably best to not use this variable in our models. Some other variables, including the final body weight of the calf represented by `Final.Calf.BW`, are missing in 19 entries. Of the other four variables the client was most interested in, none have more than ten missing values.

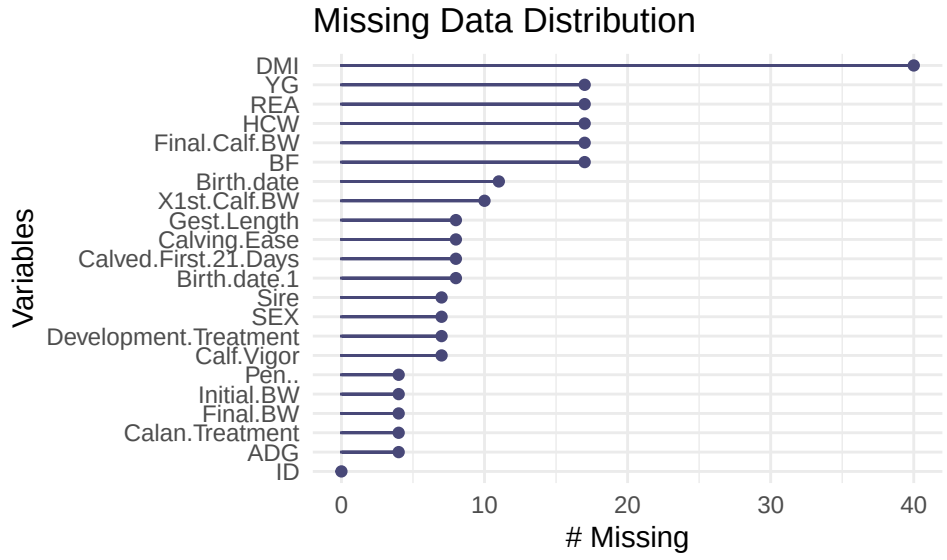


Figure 1: Chart counting the number of missing values for each variable within the data.

Cleaning the Dataset

Due to the missing values discussed above, we need to clean this dataset before continuing. There are generally two ways to handle missing values. The first is imputing them with some metric like the mean, median, or mode of the variable. The second is to just remove any rows with missing data.

We decided to cut all rows that contained missing values for variables we are interested in. These variables include the five variables the client is interested in, but also the pen number, sire, and the initial weight of the mother. The latter two are `Sire` and `Initial.BW` in the dataset. After removing missing values for these entries the dataset has 101 rows, which we feel is an ample amount for analyzing the data. Note all future figures and models come from this cleaned dataset and not the original.

We initially considered imputing the quantitative variables with the respective median values and using the mode for categorical variables as Memon, Wamala, and Kabano (2023) suggests. This has some issues though. For an example let's consider the third trimester treatment. The MET treatment was used in 40 cases, while the other two treatments were only used 38 times, meaning there are four missing values. If we mode impute this variable there will be 42 instances of the MET treatment. However, it seems very possible that the missing entries were split between the CON and DDG treatments to make an even 40 uses each. While imputing quantitative variables is less risky, we are not fully comfortable with that approach either since we are trying to analyze the data.

Summary Statistics

Let's take a closer look at what the three dependent variables the client is interested in. Figure 2 shows several summary statistics for each. Looking at the maximum values of the calving ease and calf vigor, we can see that the cleaned dataset does not contain any instances of poor scores for either. Both scores only goes from one to three. Note, the original dataset did have three instances of a vigor score of four or five, but each row was missing a final body weight so the entries were not included in the cleaned dataset. We can also see from the median and 75th percentiles that both seem very skewed towards the low end of the scale. While this is a good thing in terms of the health of the cows it could present some challenges for us later on.

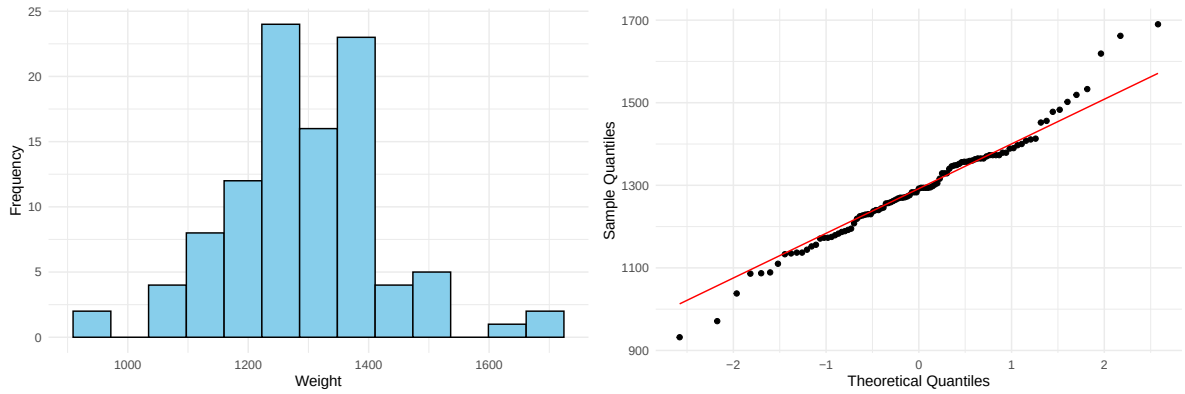
The final weight of calf is the third variable in Figure 2. The mean and median are relatively similar given the large standard deviation. While the previous two variables discussed give us some concerns about the skew, the final weight does not present the same issues. Further investigation into the approximate distribution of the final weight is needed though.

Let's look at a histogram and a Q-Q plot for the final weight of the calves in Figure 3. The bin width for the histogram comes from the Freedman-Diaconis rule. (William (2023)). The

Variable	Mean	Median	SD	25th Percentile	75th Percentile	Min	Max
Calving.Ease	1.118812	1	0.4071149	1	1	1	3
Calf.Vigor	1.257426	1	0.6107940	1	1	1	3
Final.Calf.BW	1291.584158	1292	128.9883148	1219	1365	932	1690

Figure 2: Summary statistics for dependent variables

histogram appears to follow an approximately normal distribution. The Q-Q plot mostly follows this as most points follow the linear trend represented by the red line.

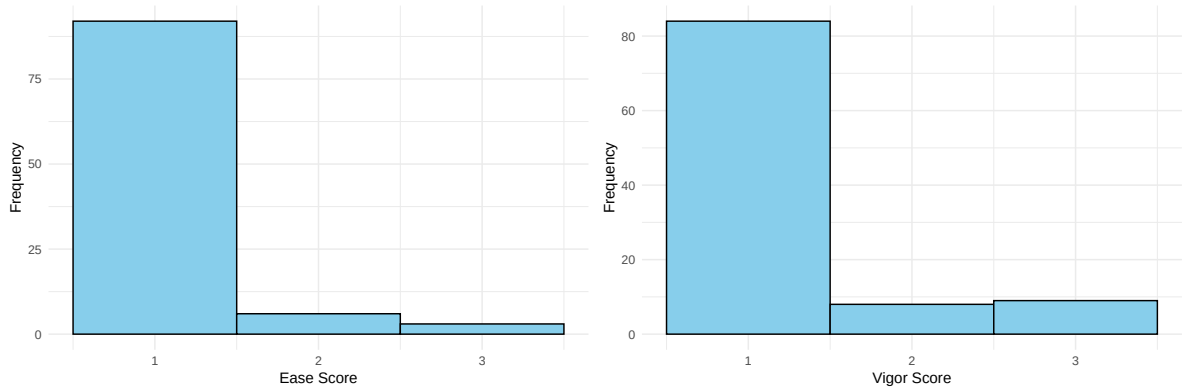


(a) Histogram of final calf weight.

(b) Q-Q plot of final calf weight.

Figure 3: Plots used to check if the distribution of the final calf weight is normal.

Before moving on we want to look at plots of the scoring variables as well. While we suspect a heavy skew for each, the histograms in Figure 4 verify this. It is important to remember that these two variables are not continuous like the weight variable, so the types of models used will vary.



(a) Histogram of calving ease.

(b) Histogram of calf vigor.

Figure 4: Histograms of scoring variables.

Exploring the Data

Before looking at potential models, let's explore how some of the variables interact with each other. While we will be able to include other explanatory variables, the client specifically mentions using the third trimester treatment and the sex of a calf as explanatory variables of interest. The table in Figure 5 shows the breakdown of treatment by sex. Note HFR stands for heifer and STR stands for steer. Although not every group has an equivalent number of subjects, this is nothing we are concerned about. Please note that the total occurrences per treatment are different than discussed above since rows containing missing values were removed.

	HFR	STR
CON	19	15
DDG	14	20
MET	18	15

Figure 5: Table showing the breakdown of treatment by sex.

One of the key assumptions for some of the models we will be discussing later is that the explanatory variables of the model are not highly correlated with each other. If this assumption is violated, multicollinearity is present. Since both the treatment and the sex of the calf are categorical, we can use the Pearson's chi-squared on the table in Figure 5 to determine if multicollinearity is a problem for these variables. (Bhalla (2017)). The results of the test, shown in Figure 6, indicate multicollinearity is not a problem since the p-value is well above any commonly used significance level such as 0.05.

Metric	Value
Statistic	1.79
Degrees of Freedom	2
P-Value	0.40811

Figure 6: Chi-squared test for treatment and sex.

Now let's consider how these variables affect the final body weight. The boxplot shown in Figure 7 allows us to see this relationship graphically. We can see the steers are heavier on average than the heifers. The treatments seem to different variances as well, but their median values are not different by huge quantities. Both the CON and MET treatments had one steer large enough to be an outlier, while the DDG treatment had several outliers for both sexes.

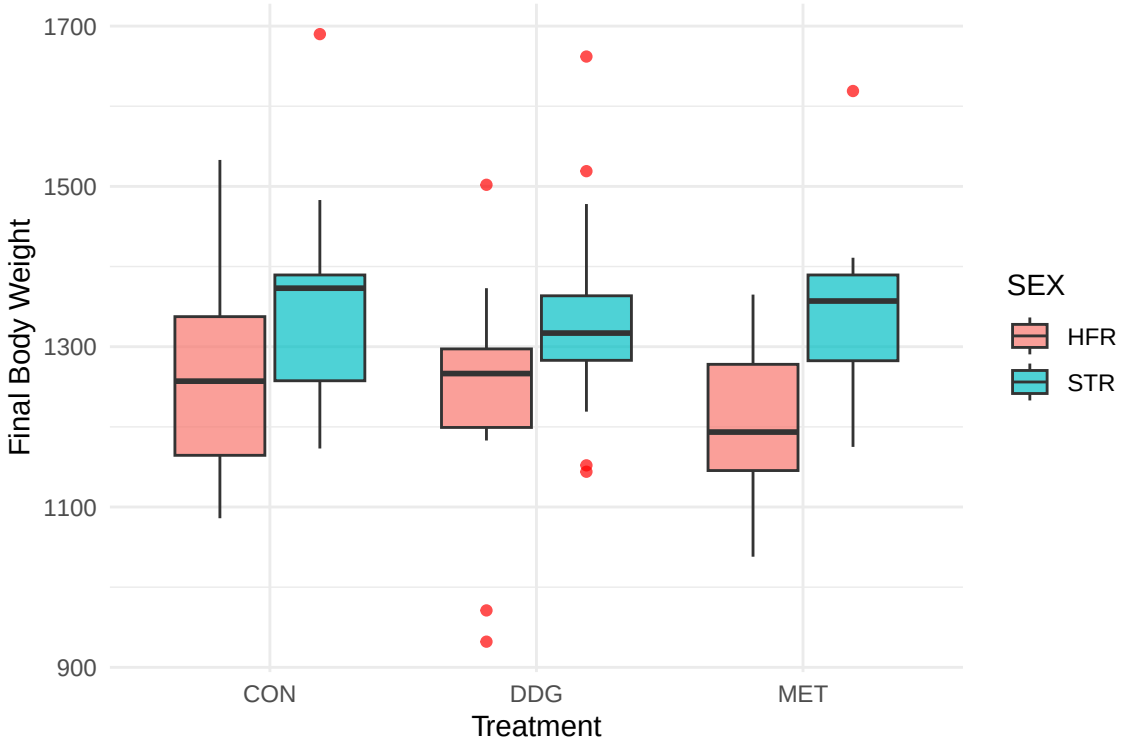
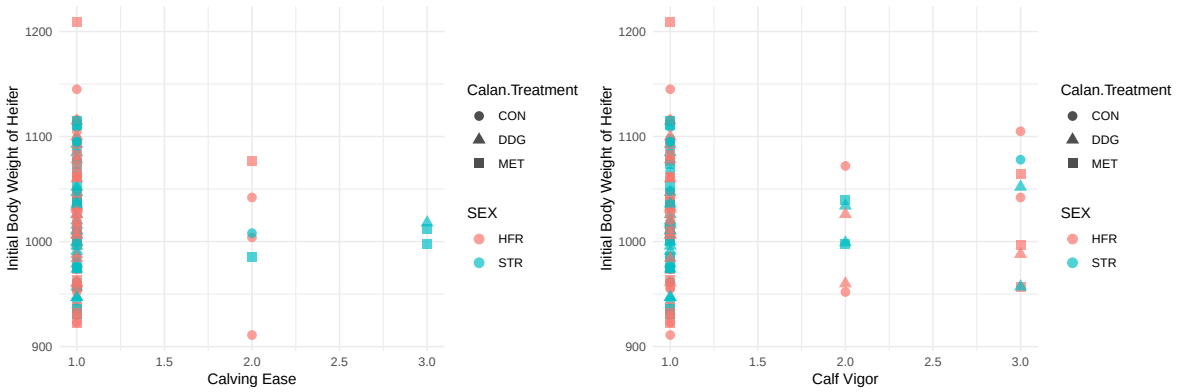


Figure 7: Boxplot of final body weight by treatment and sex.

Another variable we want to investigate graphically is the initial weight of the heifer that birthed the calf and see how it compares to both the ease and vigor score. In Figure 8 we can see this while also accounting for both the third trimester treatment with the shape of the

data point and the sex of the calf with the color of the data point. This allows us to see both how the effect the heifer's initial weight has, but also the trends of both treatment and sex.



(a) Initial body weight vs. calf vigor.

(b) Initial body weight vs. calving ease.

Figure 8: Scatterplot of heifer's initial body weight versus scoring variables, controlling for third trimester treatment and sex of the calf.

In both plots, we can see the initial weight of the mother heifer does not seem to have a huge effect on either score. It would appear that being heavier or lighter than average had little to no effect on getting a score other than one. While this variable may be used in models later to test its significance more formally, we have doubts about its effect on either score variable.

Models for Calving Ease Score

Ordinal Logistic Regression Model

Since calving ease is a score from one to three in our dataset and one to five in real life, it can be considered an ordinal variable. This means instead of treating it as a quantitative variable as we did previously, it could be considered an ordered categorical variable. One method of modeling ordinal variables is with ordinal logistic regression. This uses one or more independent variables to predict the ordinal value of the dependent variable. A key assumption of ordinal logistic regression, outside of the dependent variable being ordinal, is no multicollinearity between independent variables. ("Ordinal Regression" (n.d.)).

Let's attempt to apply an ordinal logistic regression to the calving ease variable. In this simple case we will use the third trimester treatment and the calf's sex as the independent variables. As we saw in Figure 6, multicollinearity is not a problem with these variables, so we may proceed.

Ordinal logistic regression models are expressed in terms of a logit function containing probabilities rather than a single variable. In this case, where Y is a random variable representing the calving ease score and $l = 1, 2, ..., 5$ represents the possible values of Y in theory and $l = 1, 2, 3$ in our dataset, $\frac{P(Y \leq l)}{P(Y > l)}$ is the cumulative probability of the easing score being less than or equal to level l versus greater than l. We can then write the model as

$$\ln\left(\frac{P(Y \leq l)}{P(Y > l)}\right) = \alpha_l - \beta_{trt}X_{trt} - \beta_{sex}X_{sex} - \beta_{trt:sex}(X_{trt} \times X_{sex})$$

In the equation above the left side is the log-odds or log of the cumulative probability, which is the logit function mentioned previously. On the right side of the equation α_l is the intercept of the model at the lth level, X_{trt} and X_{sex} represent the values for the third trimester treatment and the sex respectively, and the β_i are slope values corresponding to the ith explanatory variable with $\beta_{trt:sex}$ being the interaction effect.

Some of the results of the model are displayed in Figure 9. Note the model treats CON and heifer as the starting values for the treatment and sex respectively. The first three rows of the coefficients table report the main effects for the other two treatments and steers. `Calan.TreatmentDDG:SEXST` respresnts the interaction effect between the DDG treatment and steers, which is highly significant (p-value < 0.001). The interaction bewtween the MET treatment and steers is not significant though (p-value $= 0.149$). Overall, the model highlights significant effects for DDG and its interaction with calf sex, while other predictors do not show substantial influence.

	Value	Std. Error	t value	p-value
Calan.TreatmentDDG	-16.863	0.728	-23.147	0.000
Calan.TreatmentMET	-1.116	1.203	-0.928	0.353
SEXSTR	-0.926	1.208	-0.767	0.443
Calan.TreatmentDDG:SEXSTR	16.616	0.728	22.809	0.000
Calan.TreatmentMET:SEXSTR	2.473	1.714	1.443	0.149
1|2	1.735	0.625	2.776	0.006
2|3	2.939	0.781	3.763	0.000

Figure 9: Coefficients table for ordinal logistic regression.

The last two rows of Figure 9 provide estimates for the interepts of the model. For instance the row `1|2` means for $l <= 1$ the model estimates $\alpha_1 = 1.735$ and for $l <= 2$ we can say $\alpha_2 = 2.939$. Both of these interecpts are significant as well.

In summary of this model, we found a substantial difference between heifer calves from mothers on the CON treatment compared to steer calves with mothers on the DDG treatment as it relates to the easing score. The DDG steer interaction effect had an estimated positive

coeffienct, but in the context of the model this implies a decrease in the log-odds. The MET treatment did not provide a statistically significant effect though. We did revcieve and AIC of 79.95121 as well, which seemed decent to us. ("Ordinal Logistic Regression | r Data Analysis Examples" (2011)).

Multinomial Logistic Regression

In search of another model we considered a multinomial logistic regression. The model used the same independent variables as before, but resulted in fewer significant terms. While the AIC of this model indicated it still fit decently, we are not showing output for this model due to concerns with the nature of the data. Multinomial logistic regression is designed to model nominal data, which is categorical data without an order. (Frost (2021)). For these reasons we chose to go in another direction and not recommend this model.

Binomial Regression

As previously discussed, the data for the calving ease score is highly skewed with most calves getting a score of one. One way to reduce the effect of the skew could be to convert this to a binomial data set. In this case we would have two categories for the calving ease score: low, which corresponds to scores of one, and high, which is everything else. A type of model that fits binomial data like this is a binomial regression. While this may not make a huge difference given that we have no scores larger than three, we want to see how this model compares to the last.

For this model, we chose to use the same two explanatory variables, third trimester treatment and sex of the calf. In this case let π be the probability of a high score given the explanatory variables, X_{trt} and X_{sex} respectively. In other words $\pi = P(Y = high|X_{trt}, X_{sex})$ where Y is the same random variable as before, but is only ever *low* or *high*. Binomial regression uses π in a similar logit function as the ordinal logistic regression. Thus we can write the model as,

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_{trt} + \beta_2 X_{sex} + \beta_3 (X_{trt} \times X_{sex})$$

where β_0 is the intercept of the model and β_1, β_2 and β_3 are the coefficients for X_{trt}, X_{sex}, and their interaction respectively. (Wiley (2013)). For this model MET is considered to be the default treatment, while steer is the default sex.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	62.695	66.599
SC	65.310	82.290
-2 Log L	60.695	54.599

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.0963	5	0.2970
Score	5.4692	5	0.3613
Wald	3.0592	5	0.6909

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Calan.Treatment	2	2.1324	0.3443
SEX	1	1.4189	0.2336
Calan.Treatment*SEX	2	1.9791	0.3718

Figure 10: *Fit Statistics* and other tests.

In the *Joint Tests* table above we can see the interaction terms are insignificant. This means we can look at main effects for each variable level. These can be seen in the *Analysis of Maximum Likelihood Estimates* below, but all terms outside of the intercept are insignificant as well.

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			1	-1.3863	0.6455	4.6123	0.0317
Calan.Treatment	CON		1	-1.2528	1.2199	1.0546	0.3044
Calan.Treatment	DDG		1	-1.5581	1.2121	1.6524	0.1986
SEX	HFR		1	-1.4469	1.2147	1.4189	0.2336
Calan.Treatment*SEX	CON	HFR	1	2.4120	1.7154	1.9770	0.1597
Calan.Treatment*SEX	DDG	HFR	1	-9.0820	225.2	0.0016	0.9678

Figure 11: *Analysis of Maximum Likelihood Estimates* table.

Elsewhere in the output, which can be seen in *Appendix C - Additional SAS Output*, the odds ratios and subsequent confidence intervals also indicated no significant results.

While the lack of significant terms seems detrimental, the AIC of the model is 66.599, indicating a decent fit. However, we cannot compare this value to the AIC of the ordinal logistic model since they are different types of models. (Hyndman (2013)). This is backed up by the percent concordance (63.9%) and Somers' D (0.440), which suggest moderate predictive performance.

Recommended Model/Conclusion for Calving Ease Score

We previously discussed three models for the calving ease score. As we already indicated, the multinomial logistic regression is not the direction we would recommend. In regards to the other two, model preference depends on the cleint's desired goal. If these models are intended for purely predictive purposes, the binomial regression model may be effective. However, if the client wants to study the impact the third trimester treatment and sex of the calf had on the ease score then we strongly recommend the ordinal logistic regression model over the other models discussed.

Models for Calf Vigor Score

Ordinal Logistic Regression Model

The similarities between the calving ease and calf vigor score make identifying potential models a much easier task, as what might fit calving ease score can also potentially fit the calf vigor

score. We will first look at an ordinal logistic regression model. The model can once again be represented by the equation

$$\ln\left(\frac{P(Y \leq l)}{P(Y > l)}\right) = \alpha_l - \beta_{trt}X_{trt} - \beta_{sex}X_{sex} - \beta_{trt:sex}(X_{trt} \times X_{sex})$$

since the client is interested in the effect of the same explanatory variables on the score. This equation looks exactly the same as before, with the only changes being Y now representing the calf vigor score and l now ranging from one to eleven in real life. (Probo (2022)). Note, within out cleaned data set, $l = 1, 2, 3$ as it did for the ease score.

While the dependent variable and equation are very similar to what we previously did the results of the model are not. In Figure 12 we can see the interaction terms are non-significant. The same can be said for the main effects meaning only the intercepts, α_l, are significant. This indicates neither of the explanatory variables seems to have much of an effect on the vigor score.

	Value	Std. Error	t value	p-value
Calan.TreatmentDDG	-0.021	0.852	-0.025	0.980
Calan.TreatmentMET	-0.179	0.845	-0.211	0.833
SEXSTR	-1.271	1.177	-1.080	0.280
Calan.TreatmentDDG:SEXSTR	1.229	1.452	0.847	0.397
Calan.TreatmentMET:SEXSTR	0.834	1.535	0.543	0.587
1|2	1.332	0.559	2.381	0.017
2|3	2.068	0.601	3.440	0.001

Figure 12: Coefficients table for ordinal logistic regression.

Additionally, the model returned an AIC of 127.059. This is not bad on its own, and indicates the model fits decently. The issues regarding the significance makes us skeptical of the model's predictive capability. However, it may indicate the two explanatory variables just do not impact the vigor score in a significant way. We will want to test other models before arriving at that conclusion though.

Binomial Regression

Given the lack of significance with the ordinal logistic regression model, we will look into the binomial example again. As before we are condensing the vigor score into a *low* category and *high* category. All of the scores of one fit in the former and the rest are in the latter. Just as with the calving ease score, let π be the probability of a high score. Then, by using the same two explanatory variables again, we can write the model as

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_{trt} + \beta_2 X_{sex} + \beta_3 (X_{trt} \times X_{sex})$$

Once again, the *Analysis of Maximum Likelihood Estimates* table below indicates no significant interactions or main effects, outside of the intercept. Again though, the *Model Fit Statistics* table has a decent AIC store, so the model may still have predictive qualities despite showing no significant relationships among our explanatory variables.

Analysis of Maximum Likelihood Estimates

Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			1	1.8718	0.7596	6.0730	0.0137
Calan.Treatment	CON		1	0.7666	1.2837	0.3567	0.5504
Calan.Treatment	DDG		1	-0.4855	0.9431	0.2650	0.6067
SEX	HFR		1	-0.2624	0.9884	0.0705	0.7907
Calan.Treatment*SEX	CON	HFR	1	-1.0543	1.5377	0.4701	0.4929
Calan.Treatment*SEX	DDG	HFR	1	0.1754	1.3091	0.0179	0.8934

Figure 13: *Analysis of Maximum Likelihood Estimates* table.

More of the SAS output, including the reported fit statistics, joint tests, and odds ratios can be found in *Appendix C - Additional SAS Output*.

Recommended Model/Conclusion for Calf Vigor Score

The two models discussed above both had a similar conclusion, which is the vigor score was not significantly affected by either the third trimester treatment, the cal's sex, or their interaction. Therefore we cannot recommend a model for this dataset. However, both models could still have some predictive value, which could be evaluated at a later date.

Note we did fit a multinomial logistic regression model for the vigor score as well, but we ran into the same issues as before so no results are shown for the model in this paper.

Models for Final Calf Body Weight

Mixed Model

Now we will look at models for investigating the effect the clients chosen explanatory variables have on the final body weight of the calf. Since the body weight is a quantitative variable, we

can turn to linear models. We initially attempted to fit a simple model using only the third trimester treatment and sex of the calf, but ultimately chose to include the sire of the calf as an additional explanatory variable. This slightly complicates the model as we believe the additional variable is better represented as a random effect, meaning we are working with a mixed model.

The purposed model and be represented as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_k + e_{ijk}$$

where Y_{ijk} represents the final weight for the ith treatment,the jth sex, and the kth sire. μ represents the overall mean, while α_i is the fixed effect for the ith treatment, β_j is the fixed effect for the jth sex, and $(\alpha\beta)_{ij}$ is their interaction. The random sire effect is represented by u_k and we assume u_k $N(0, \sigma_k^2)$. We also assume the residual term e_{ijkl} is distributed as $N(0, \sigma^2)$.

For a linear mixed model to work, the assumption regarding the distribution of the residuals must hold. These can be checked graphically in the conditional residual plots below. The histogram and Q-Q plot show if the residuals are approximately normal while the plot in the top left evaluates any multicollinearity concerns. Thankfully the graphs give us no concerns as no trends are present in the top-left plot and the residuals are randomly distributed around zero, the histogram appears bell shaped around zero, and the Q-Q plot sees most points fall along the line, which is ideal. Therefore we can procede with our mixed model.

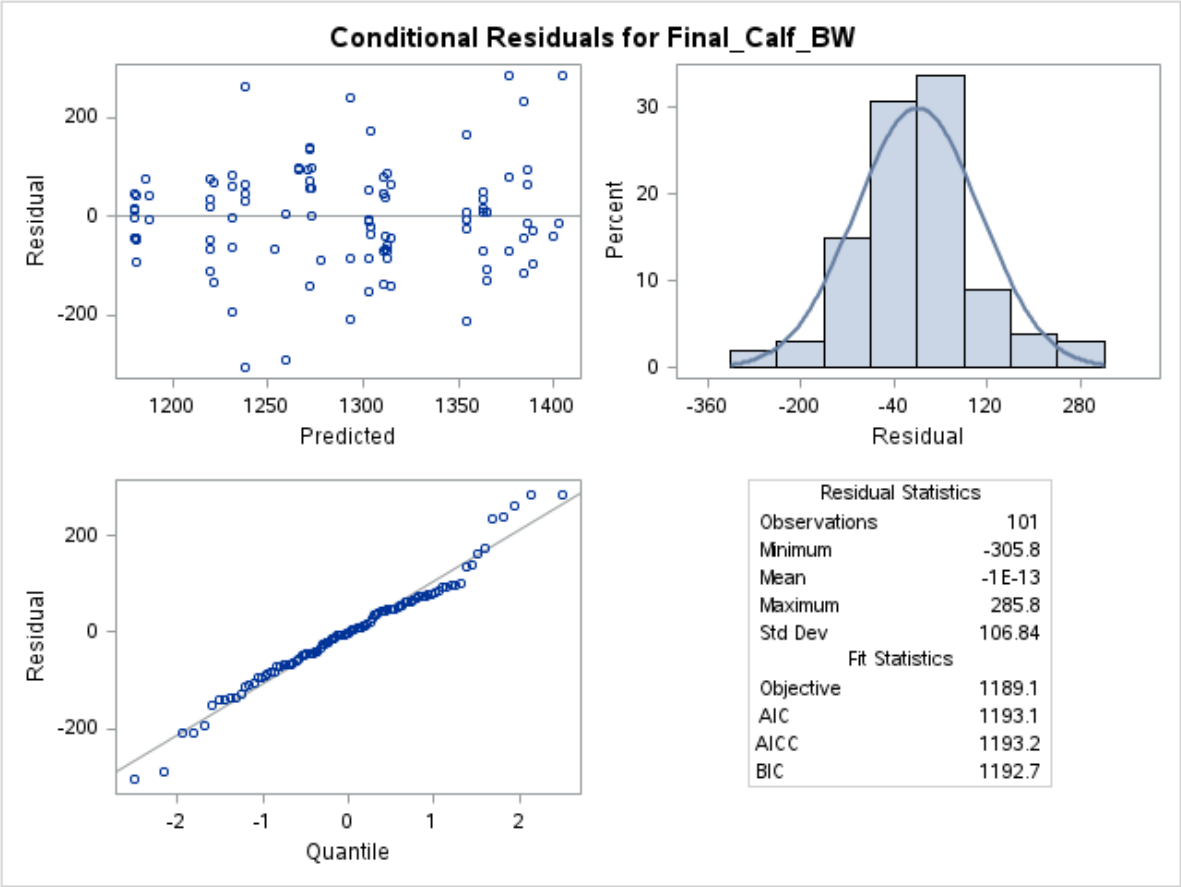


Figure 14: Plots to check residual assumption.

The *Covariance Parameter Estimates* table shows how the sire effect

From the *Fit Statistics* table below shows the model has an AIC of 1193.1. While this is not great, it is not far off from other models we have viewed in other data analyses. Also, the sire effect on the covariance structure is shown in the *Covariance Parameter Estimates* table.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	Sire	2372.00
Residual		12484

Fit Statistics	
-2 Res Log Likelihood	1189.1
AIC (Smaller is Better)	1193.1
AICC (Smaller is Better)	1193.2
BIC (Smaller is Better)	1192.7

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Calan.Treatment	2	90	0.41	0.6660
SEX	1	90	25.16	<.0001
Calan.Treatment*SEX	2	90	0.23	0.7919

Figure 15: Covariance parameter, fit statistics, fixed effects and least quares means tables

The *Type 3 Fixed Effects* table tells us what fixed effects, if any, are statistically significant. As we have done when viewing other models in this paper, we must first look at the interaction effect, which is insignificant. Therefore, we turn to the main effects of the two explanatory variables. While the sex of the calf is highly significant (p-value < 0.0001), the treatment effect is insignificant. Given that the interaction is insignificant, we chose not to evaluate the *Differences of Least Squares Means* table, but we did include it in *Appendix C - Additional SAS Output* for the sake of transparency.

ANCOVA Model

Another model that could fit the data is an ANCOVA model. These models take a continuous dependent variable and takes at least one categorical variable along with a covariate. For the dependent and categorical variables we will once again be the third trimester treatment and sex of the calf. The covariate must be a continuous, independent variable that is not a primary interest of the study. (Frost (2023)). We previously looked at the mother heifer's initial body weight in Figure 8 so we will make that the covariate.

Now we can view the output of this model after it was fit in SAS. The first table in the figure below shows the sum of squares for the model. By dividing the model sum of square by the corrected total sum of squares we can say the model accounts for approximately 26.75% of the variation in the final body weight variable.

The GLM Procedure

Dependent Variable: Final_Calf_BW

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	444984.875	74164.146	5.72	<.0001
Error	94	1218813.659	12966.103		
Corrected Total	100	1663798.535			

R-Square	Coeff Var	Root MSE	Final_Calf_BW Mean
0.267451	8.816212	113.8688	1291.584

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Calan.Treatment	2	13661.9170	6830.9585	0.53	0.5922
SEX	1	294400.0151	294400.0151	22.71	<.0001
Calan.Treatment*SEX	2	8623.2008	4311.6004	0.33	0.7179
Initial BW	1	128299.7422	128299.7422	9.90	0.0022

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Calan.Treatment	2	8383.2348	4191.6174	0.32	0.7246
SEX	1	284570.2895	284570.2895	21.95	<.0001
Calan.Treatment*SEX	2	16304.4068	8152.2034	0.63	0.5355
Initial BW	1	128299.7422	128299.7422	9.90	0.0022

Figure 16: Model information along with Type I and Type III test results.

The last two tables show the results of the Type I and Type III sums of squares tables respectively. We are more interested in the Type III sums of squares in this case. Once again though, the interaction is insignificant and out of the two categorical explanatory variables only the sex of the calf has a significant main effect (p-value < 0.0001). It should be noted that the effect of the initial body weight of the mother is significant as well (p-value=0.0022).

I NEED HELP INTERPRETING GRAPHS

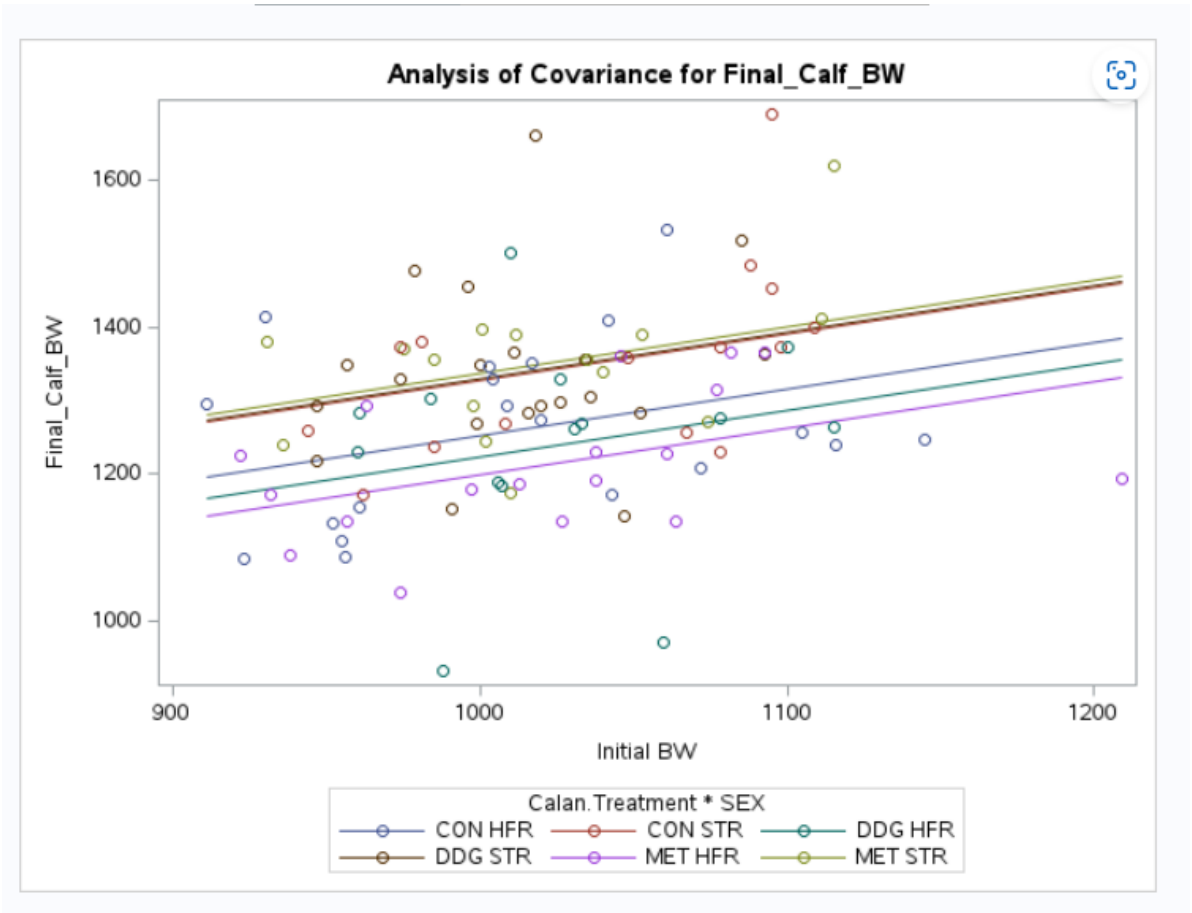


Figure 17: Fig-1

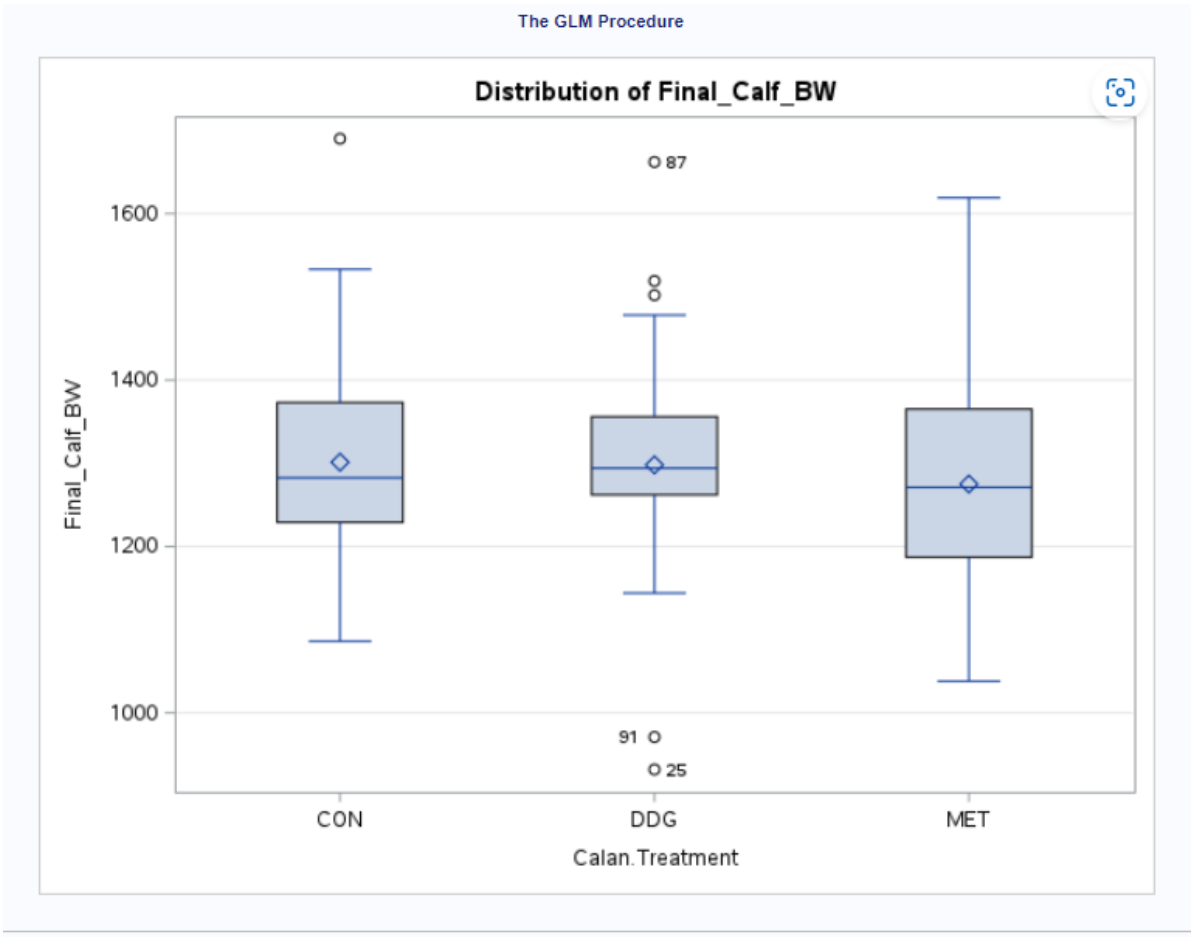


Figure 18: Fig-1

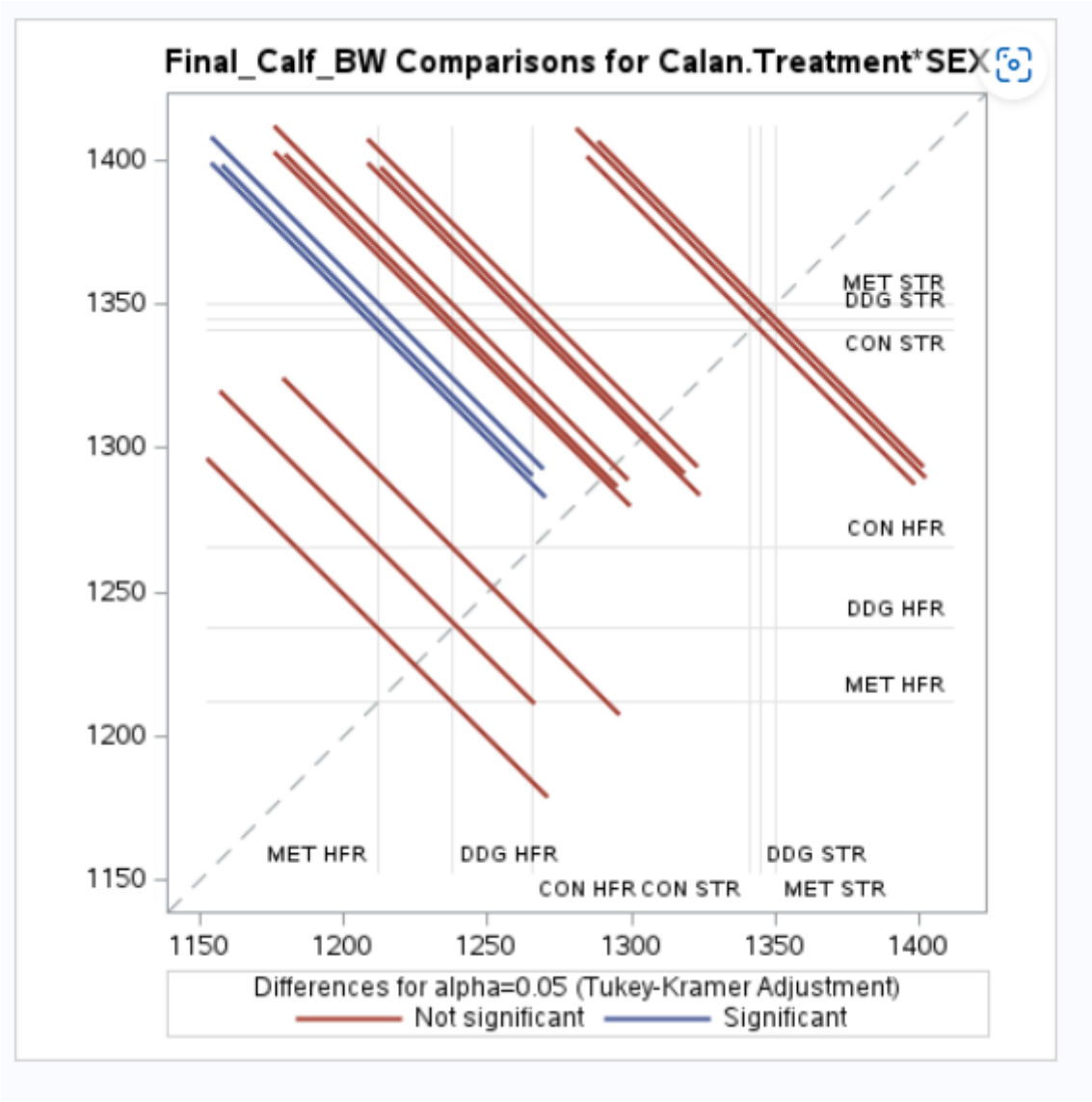


Figure 19: Fig-1

Checking Assumptions

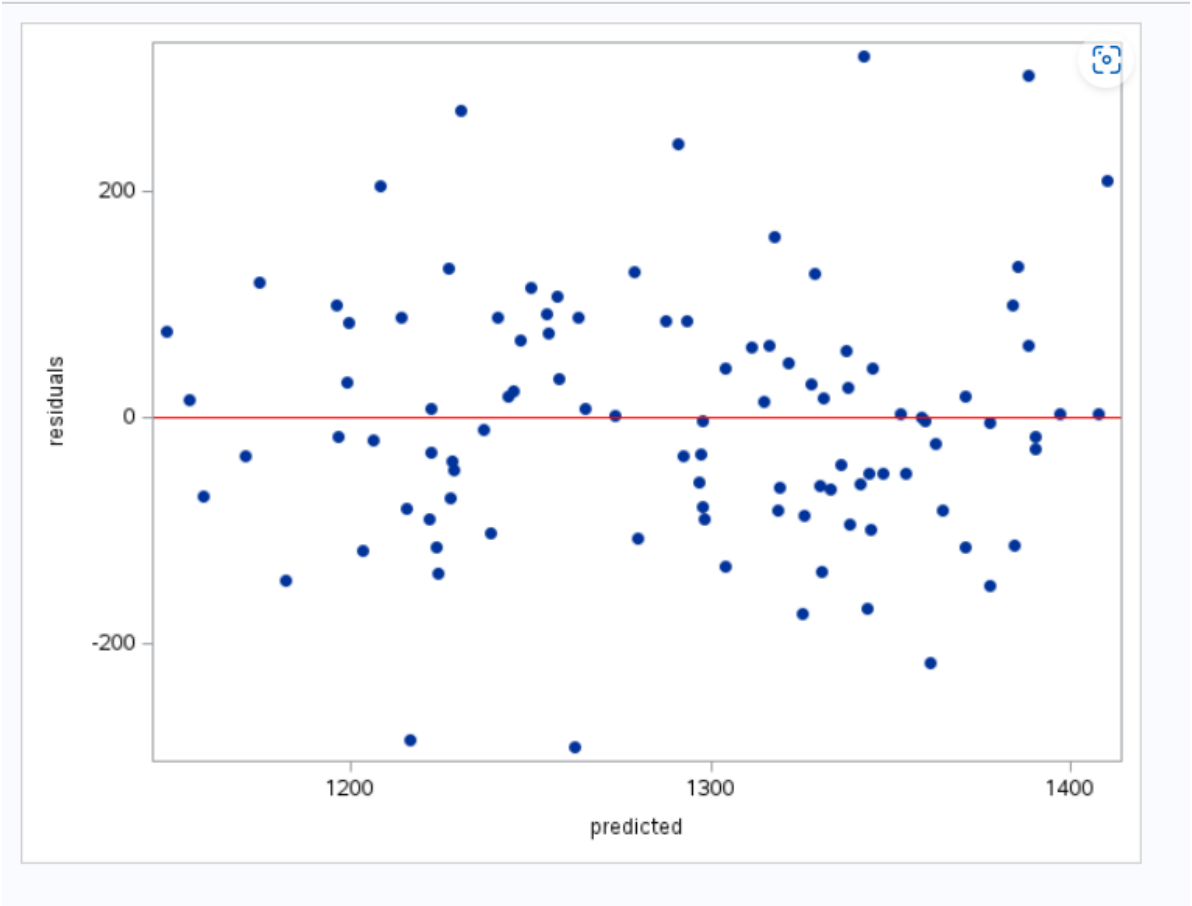


Figure 20: Fig-1

The UNIVARIATE Procedure
Variable: residuals

Moments			
N	101	Sum Weights	101
Mean	0	Sum Observations	0
Std Deviation	110.399894	Variance	12188.1366
Skewness	0.30694177	Kurtosis	0.87710048
Uncorrected SS	1218813.66	Corrected SS	1218813.66
Coeff Variation	.	Std Error Mean	10.9852

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	110.39989
Median	-2.30039	Variance	12188
Mode	.	Range	610.59481
		Interquartile Range	140.56531

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-0.5	Pr >= |M|	1.0000
Signed Rank	S	-92.5	Pr >= |S|	0.7557

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.981261	Pr < W	0.1618
Kolmogorov-Smirnov	D	0.054393	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.057752	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.48068	Pr > A-Sq	0.2342

Figure 21: Fig-1

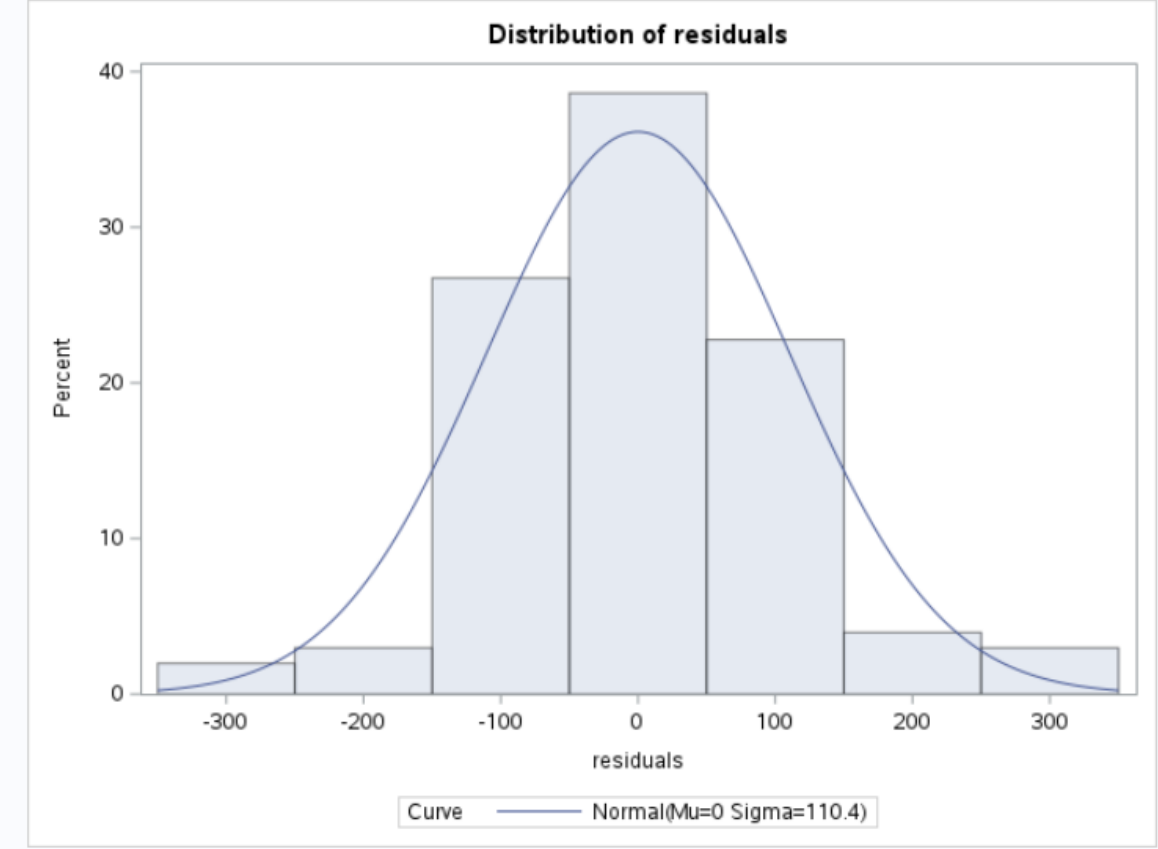


Figure 22: Fig-1

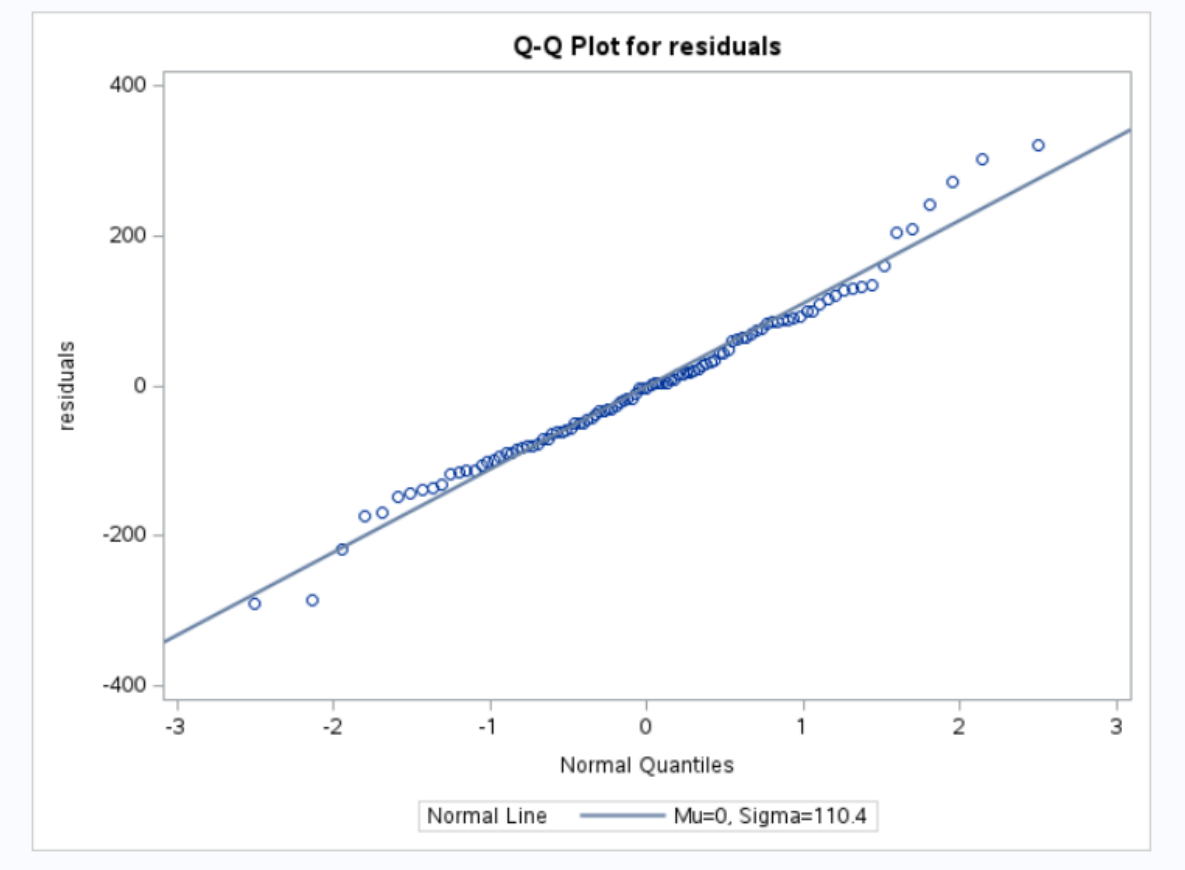


Figure 23: Fig-1

Recommended Model/Conclusion for Final Calf Body Weight

From the results of both models we can say the sex of the calf had a significant impact on the final body weight of the calf. There was not a significant result for the third trimester treatment effects though.

It is important to note that in the ANCOVA model, the initial body weight of the mother was deemed significant. In the future including more models that account for this variable and others within the data set may give cattle producers a better idea of how to maximize the weight of their calves before slaughter.

Conclusion

This dataset contained many missing values. We ended up performing our analysis on only 84.17% of the original data. Repeating our analysis methods with imputed data would allow for more of the data to be used and may lead to slightly different results.

Additionally, future analysis of this data should attempt to control for more variables in the models. In our final model, the ANCOVA model for the final body weight of the calves, we found information about the mother heifer, her body weight, had a significant impact on the model. We were unable to apply this variable and others to additional models due to time constraints, but would recommend this to anyone working on this data in the future.

References

Bhalla, Deepanshu. 2017. "How to Check Multicollinearity in Categorical Variables." Listening Data. 2017. https://www.listendata.com/2015/04/detecting-multicollinearity-in-categorical-variables.html#:~:text=When%20dealing%20with%20nominal%20variables, suggest%20the%20presence%20of%20multicollinearity.

Frost, Jim. 2021. "Multinomial Logistic Regression: Overview & Example." Statistics By Jim. 2021. https://statisticsbyjim.com/regression/multinomial-logistic-regression/.

———. 2023. "ANCOVA: Uses, Assumptions & Example." Statistics By Jim. 2023. https://statisticsbyjim.com/anova/ancova/#comments.

Heins, Glenda, Brad & Pereira. 2023. "Monitoring Calving Traits to Improve Cow and Calf Health." University of Minnesota Extension. 2023. https://extension.umn.edu/dairy-milking-cows/calving-traits#:~:text=Calving%20ease%20score%201%3A%20quick, score%204%3A%20used%20obstetrical%20chains.

Hyndman, Ron J. 2013. "Facts and Fallacies of the AIC." 2013. https://robjhyndman.com/hyndsight/aic/.

Memon, Shaheen MZ., Robert Wamala, and Ignace H. Kabano. 2023. "A Comparison of Imputation Methods for Categorical Data." *Informatics in Medicine Unlocked* 42: 101382. https://doi.org/https://doi.org/10.1016/j.imu.2023.101382.

"Ordinal Logistic Regression | r Data Analysis Examples." 2011. UCLA. 2011. https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/.

"Ordinal Regression." n.d. University of St. Andrews. n.d. https://www.st-andrews.ac.uk/media/ceed/students/mathssupport/ordinal%20logistic%20regression.pdf.

Probo, Maria Cristina, Monica & Veronesi. 2022. "Clinical Scoring Systems in the Newborn Calf: An Overview." *Animals (Basel)*. https://doi.org/doi: 10.3390/ani12213013.

Saner, Brianna, Randy & Buseman. 2024. "How Many Pounds of Meat Can We Expect from a Beef Animal?" 2024. https://beef.unl.edu/beefwatch/2020/how-many-pounds-meat-can-we-expect-beef-animal.

USDA. "5017-1: Calculating Dry Matter Intake from Pasture." https://www.ams.usda.gov/rules-regulations/organic/handbook/5017-1#:~:text=DMI%20is%20the%20level%20of, life%20and%20level%20of%20production.

Wiley, John. 2013. "Binomial and Poisson Regression." 2013. https://www2.stat.duke.edu/courses/Fall21/sta521.001/post/week08-2/Applied_Linear_Regression_----_%28CHAPTER_12%29.pdf.

William, Nkugwa Mark. 2023. "How to Determine Bin Width for a Histogram (r and Python)." 2023. https://nkugwamarkwilliam.medium.com/how-to-determine-bin-width-for-a-histogram-r-and-pyth-653598ab0d1c.

Appendix A - R Code

```
library(knitr)
library(dplyr)
library(ggplot2)
library(naniar)
library(reshape2)
library(GGally)
library(janitor)
library(emmeans)
library(MASS)
library(multcomp)
library(lme4)
library(nnet)

data <- read.csv("data.csv")
# Placeholder of original data
origdata <- data
data <- data %>%
  mutate(across(everything(), ~ ifelse(. %in% c(".", ""), NA, .)))

length(unique(na.omit(data$Sire)))
length(unique(na.omit(data$Development.Treatment)))
sum(rowSums(is.na(data)) > 0)
rows_with_na <- sum(rowSums(is.na(data)) > 0)
unique_entries <- unique(data$Calan.Treatment)
num_unique <- length(unique_entries)
frequency_table <- table(data$Calan.Treatment)

gg_miss_var(data) +
  ggtitle("Missing Data Distribution") +
  theme_minimal()

var_used <- c("Calan.Treatment", "SEX", "Calving.Ease",
              "Calf.Vigor", "Final.Calf.BW", "Pen..", "Initial.BW", "Sire")
cleaned_data <- data[complete.cases(data[, var_used]), ]

dep_vars <- c("Calving.Ease", "Calf.Vigor", "Final.Calf.BW")

custom_names <- c("Calving Ease", "Calf Vigor", "Final Calf Weight")
```

```r
cleaned_data$Calving.Ease <- as.numeric(cleaned_data$Calving.Ease)
cleaned_data$Calf.Vigor <- as.numeric(cleaned_data$Calf.Vigor)
cleaned_data$Final.Calf.BW <- as.numeric(cleaned_data$Final.Calf.BW)

calc_stats <- function(var) {
  mean_val <- mean(var, na.rm = TRUE)
  median_val <- median(var, na.rm = TRUE)
  sd_val <- sd(var, na.rm = TRUE)
  quantiles <- quantile(var, probs = c(0.25, 0.75), na.rm = TRUE)
  max <- max(var, na.rm = TRUE)
  min <- min(var, na.rm = TRUE)
  c(mean = mean_val, median = median_val, sd = sd_val,
    Q1 = quantiles[1], Q3 = quantiles[2], Min = min, Max = max)
}

summary_table <- t(sapply(dep_vars, function(var)
  calc_stats(cleaned_data[[var]])))

summary_table <- as.data.frame(summary_table)

kable(summary_table, col.names = c("Variable", "Mean",
                                   "Median", "SD", "25th Percentile",
                                   "75th Percentile", "Min", "Max"),
      caption = "Summary Statistics for Dependent Variables")

bin_width <- round(2*IQR(cleaned_data$Final.Calf.BW)/
                     (length(cleaned_data$Final.Calf.BW))^(1/3), 2)

ggplot(cleaned_data, aes(x = Final.Calf.BW)) +
  geom_histogram(binwidth = bin_width, color = "black", fill = "skyblue") +
  labs(
    x = "Weight",
    y = "Frequency"
  ) +
  theme_minimal()

ggplot(cleaned_data, aes(sample = Final.Calf.BW)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
```

```
  ) +
  theme_minimal()

ggplot(cleaned_data, aes(x = Calving.Ease)) +
  geom_histogram(binwidth = 1, color = "black", fill = "skyblue") +
  labs(
    x = "Ease Score",
    y = "Frequency"
  )+
  theme_minimal()

ggplot(cleaned_data, aes(x = Calf.Vigor)) +
  geom_histogram(binwidth = 1, color = "black", fill = "skyblue") +
  labs(
    x = "Vigor Score",
    y = "Frequency"
  ) +
  theme_minimal()

data2 <- cleaned_data

crosstab_treatment_sex <- table(data2$Calan.Treatment, data2$SEX)

kable(
  crosstab_treatment_sex)

chi_sq_test <- chisq.test(crosstab_treatment_sex)
chi_sq_table <- data.frame(
  Metric = c("Statistic", "Degrees of Freedom", "P-Value"),
  Value = c(round(chi_sq_test$statistic, 2),
            chi_sq_test$parameter,
            format.pval(chi_sq_test$p.value)),
  stringsAsFactors = FALSE
)

kable(chi_sq_table, align = c("l", "c"), row.names = FALSE)

ggplot(data2, aes(x = Calan.Treatment, y = Final.Calf.BW, fill = SEX)) +
  geom_boxplot(outlier.color = "red", alpha = 0.7) +
  labs(title = "Boxplot of Final Calf Body Weight by Treatment and Sex",
       x = "Treatment", y = "Final Body Weight") +
  theme_minimal()
```

```r
data2$Initial.BW <- as.numeric(data2$Initial.BW)

ggplot(data2, aes(x = Calving.Ease, y = Initial.BW,
                  color = SEX, shape = Calan.Treatment)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(x = "Calving Ease", y = "Initial Body Weight of Heifer") +
  theme_minimal()

ggplot(data2, aes(x = Calf.Vigor, y = Initial.BW,
                  color = SEX, shape = Calan.Treatment)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(x = "Calf Vigor", y = "Initial Body Weight of Heifer") +
  theme_minimal()

data2$Calving.Ease <- factor(data2$Calving.Ease, ordered = TRUE)

ease_model_3 <- polr(Calving.Ease ~ Calan.Treatment * SEX,
                     data = data2, Hess = TRUE)

#summary(ease_model_3)

coefs <- coef(summary(ease_model_3))
p_values <- pnorm(abs(coefs[, "t value"]), lower.tail = FALSE) * 2
coefs <- cbind(coefs, "p-value" = p_values)

kable(coefs,
      digits = 3,
      format = "markdown")

data2$Calf.Vigor <- factor(data2$Calf.Vigor, ordered = TRUE)

vigor_model_1 <- polr(Calf.Vigor ~ Calan.Treatment * SEX, data = data2, Hess = TRUE)

summary(vigor_model_1)

coefs <- coef(summary(vigor_model_1))
p_values <- pnorm(abs(coefs[, "t value"]), lower.tail = FALSE) * 2
coefs <- cbind(coefs, "p-value" = p_values)
#print(coefs)
kable(coefs,
      digits = 3,
      format = "markdown")
```

Appendix B - SAS Code

```
/* Binomial Calving_Ease Model */
data data;
    set data;
    if 'Calving.Ease'n in (2, 3) then Binary_Ease = "High";
    else if 'Calving.Ease'n = 1 then Binary_Ease = "Low";
run;

proc freq data=data;
    tables Binary_Ease;
run;

/* Logistic regression model */
proc logistic data=data;
    class 'Calan.Treatment'n SEX / param=ref;
    model Binary_Ease(event='High') = 'Calan.Treatment'n|SEX;
    oddsratio 'Calan.Treatment'n;
    oddsratio SEX;
run;
```

```
/*Binomial Calf Vigor*/
data data;
    set data;
    /* Recode Calf.Vigor: 1 = Low, 2 and 3 = High */
    if 'Calf.Vigor'n = 1 then Binary_Vigor = "Low";
    else if 'Calf.Vigor'n in (2, 3) then Binary_Vigor = "High";
run;

proc freq data=data;
    tables Binary_Vigor;
run;
data data;
    set data;
    Binary_Vigor = strip(Binary_Vigor);
run;
data data;
    set data;
    Binary_Vigor = propcase(Binary_Vigor);
run;
```

```
proc logistic data=data descending;
    class 'Calan.Treatment'n SEX / param=ref;
    model Binary_Vigor = 'Calan.Treatment'n|SEX;
    oddsratio 'Calan.Treatment'n;
    oddsratio SEX;
run;
```

```
/*Mixed Model*/
data data;
    set data;
    if not missing('Final.Calf.BW'n) then
    Final_Calf_BW = input('Final.Calf.BW'n, best12.);
run;

proc mixed data=data method=reml plots=residualpanel;
    class 'Calan.Treatment'n SEX 'Pen..'n Sire;
    model Final_Calf_BW = 'Calan.Treatment'n|SEX ;
    random intercept / subject='Pen..'n;
    random intercept / subject=Sire;
    lsmeans 'Calan.Treatment'n*SEX / adjust=tukey pdiff;
run;
```

```
/*ANCOVA*/
proc glm data=data ;
    class 'Calan.Treatment'n SEX;
    model Final_Calf_BW = 'Calan.Treatment'n|SEX 'Initial BW'n ;
    means 'Calan.Treatment'n / tukey cldiff;
    lsmeans 'Calan.Treatment'n*SEX / adjust=tukey pdiff cl;
    output out=diagnostics r=residuals p=predicted;
run;

/* Diagnostic plots */
proc sgplot data=diagnostics;
    scatter x=predicted y=residuals / markerattrs=(symbol=circlefilled);
    refline 0 / axis=y lineattrs=(color=red);
run;

proc univariate data=diagnostics normal;
    var residuals;
    histogram residuals / normal;
    qqplot residuals / normal(mu=est sigma=est);
run;
```

Appendix C - Additional SAS Output

Binomial Regression Model for Calving Ease

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	63.9	Somers' D	0.440
Percent Discordant	19.9	Gamma	0.524
Percent Tied	16.2	Tau-a	0.072
Pairs	828	c	0.720

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
Calan.Treatment CON vs DDG at SEX=HFR	>999.999	<0.001	>999.999
Calan.Treatment CON vs MET at SEX=HFR	3.188	0.300	33.890
Calan.Treatment DDG vs MET at SEX=HFR	<0.001	<0.001	>999.999
Calan.Treatment CON vs DDG at SEX=STR	1.357	0.078	23.615
Calan.Treatment CON vs MET at SEX=STR	0.286	0.026	3.121
Calan.Treatment DDG vs MET at SEX=STR	0.211	0.020	2.265
SEX HFR vs STR at Calan.Treatment=CON	2.625	0.244	28.196
SEX HFR vs STR at Calan.Treatment=DDG	<0.001	<0.001	>999.999
SEX HFR vs STR at Calan.Treatment=MET	0.235	0.022	2.544

Figure 24: *Association of Predicted Probabilities and Observed Responses* and *Odds Ratio Estimates and Wald Confidence Intervals* tables.

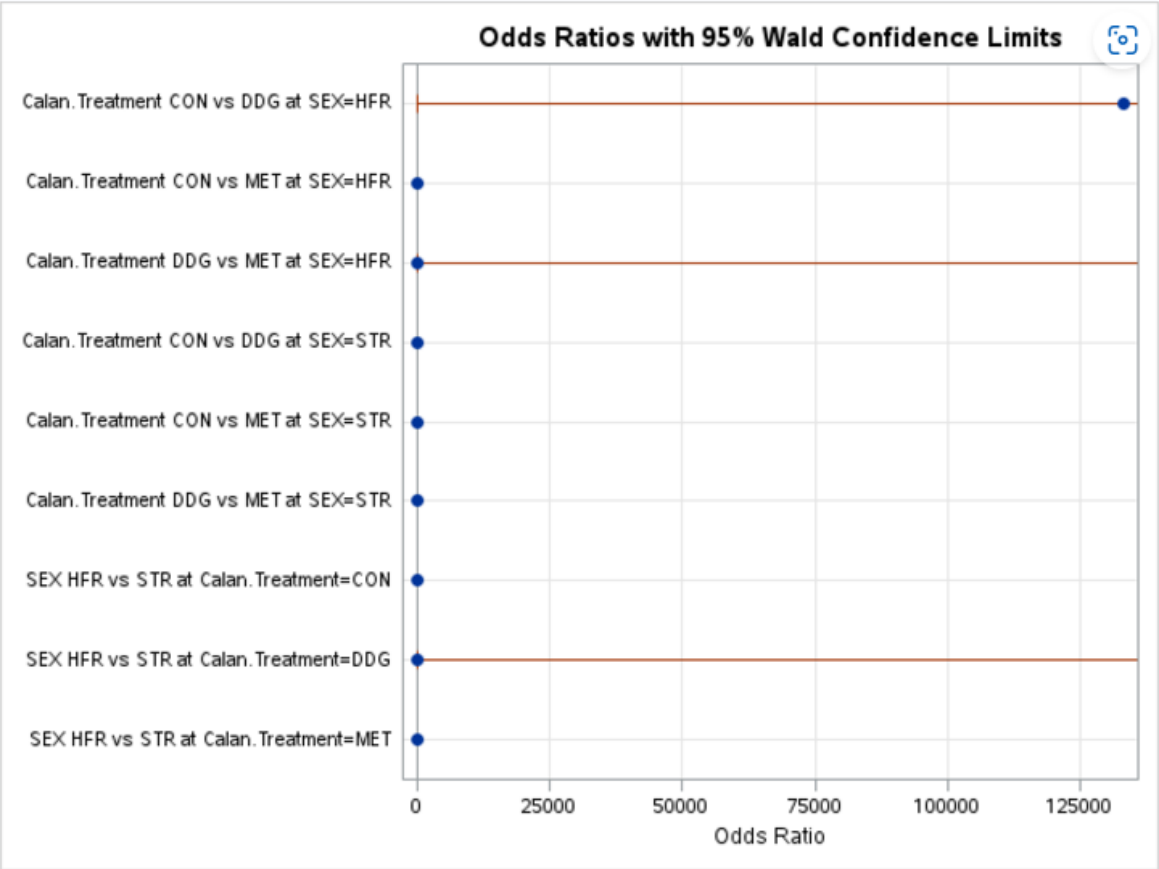


Figure 25: *Odds Ratios with 95% Wald Confidence Limits* table.

Binomial Regression Model for Calf Vigor

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	93.548	101.469
SC	96.163	117.160
-2 Log L	91.548	89.469

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.0784	5	0.8382
Score	1.8352	5	0.8714
Wald	1.6922	5	0.8899

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Calan.Treatment	2	1.1835	0.5534
SEX	1	0.0705	0.7907
Calan.Treatment*SEX	2	0.7551	0.6856

Figure 26: *Fit Statistics* and other tests.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.0	Somers' D	0.189
Percent Discordant	32.1	Gamma	0.228
Percent Tied	16.9	Tau-a	0.053
Pairs	1428	c	0.595

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
Calan.Treatment CON vs DDG at SEX=HFR	1.023	0.189	5.526
Calan.Treatment CON vs MET at SEX=HFR	0.750	0.143	3.941
Calan.Treatment DDG vs MET at SEX=HFR	0.733	0.124	4.346
Calan.Treatment CON vs DDG at SEX=STR	3.498	0.349	35.071
Calan.Treatment CON vs MET at SEX=STR	2.153	0.174	26.644
Calan.Treatment DDG vs MET at SEX=STR	0.615	0.097	3.908
SEX HFR vs STR at Calan.Treatment=CON	0.268	0.027	2.697
SEX HFR vs STR at Calan.Treatment=DDG	0.917	0.170	4.930
SEX HFR vs STR at Calan.Treatment=MET	0.769	0.111	5.338

Figure 27: *Association of Predicted Probabilities and Observed Responses* and *Odds Ratio Estimates and Wald Confidence Intervals* tables.

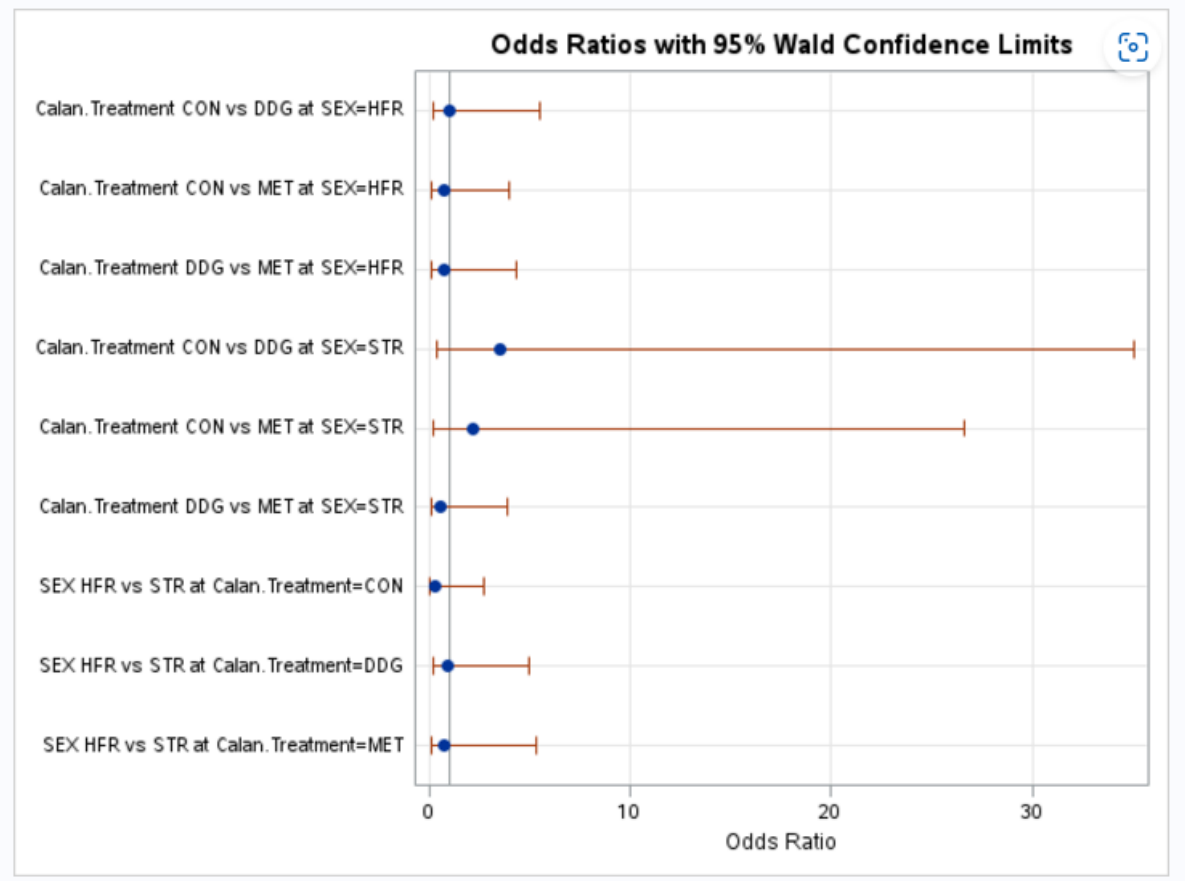


Figure 28: *Odds Ratios with 95% Wald Confidence Limits* table.

Linear Mixed Model for Calf Final Body Weight

Differences of Least Squares Means											
Effect	Calan.Treatment	SEX	_Calan.Treatment	_SEX	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
Calan.Treatment*SEX	CON	HFR	CON	STR	-93.0477	38.9261	90	-2.39	0.0189	Tukey-Kramer	0.1707
Calan.Treatment*SEX	CON	HFR	DDG	HFR	33.9699	40.4261	90	0.84	0.4030	Tukey-Kramer	0.9591
Calan.Treatment*SEX	CON	HFR	DDG	STR	-82.8198	36.4148	90	-2.27	0.0253	Tukey-Kramer	0.2155
Calan.Treatment*SEX	CON	HFR	MET	HFR	40.0888	37.2865	90	1.08	0.2852	Tukey-Kramer	0.8900
Calan.Treatment*SEX	CON	HFR	MET	STR	-90.7112	38.7579	90	-2.34	0.0215	Tukey-Kramer	0.1890
Calan.Treatment*SEX	CON	STR	DDG	HFR	127.02	41.9550	90	3.03	0.0032	Tukey-Kramer	0.0367
Calan.Treatment*SEX	CON	STR	DDG	STR	10.2279	38.2801	90	0.27	0.7899	Tukey-Kramer	0.9998
Calan.Treatment*SEX	CON	STR	MET	HFR	133.14	39.2920	90	3.39	0.0010	Tukey-Kramer	0.0130
Calan.Treatment*SEX	CON	STR	MET	STR	2.3365	41.1044	90	0.06	0.9548	Tukey-Kramer	1.0000
Calan.Treatment*SEX	DDG	HFR	DDG	STR	-116.79	39.3930	90	-2.96	0.0039	Tukey-Kramer	0.0435
Calan.Treatment*SEX	DDG	HFR	MET	HFR	6.1189	40.3190	90	0.15	0.8797	Tukey-Kramer	1.0000
Calan.Treatment*SEX	DDG	HFR	MET	STR	-124.68	42.5172	90	-2.93	0.0043	Tukey-Kramer	0.0474
Calan.Treatment*SEX	DDG	STR	MET	HFR	122.91	36.4353	90	3.37	0.0011	Tukey-Kramer	0.0136
Calan.Treatment*SEX	DDG	STR	MET	STR	-7.8914	38.5694	90	-0.20	0.8383	Tukey-Kramer	0.9999
Calan.Treatment*SEX	MET	HFR	MET	STR	-130.80	39.6146	90	-3.30	0.0014	Tukey-Kramer	0.0168

Figure 29: *Differences of Least Squares Means* table.

ANCOVA for Calf Final Body Weight

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Calan.Treatment	SEX	Final_Calf_BW LSMEAN	LSMEAN Number
CON	HFR	1265.70753	1
CON	STR	1341.33474	2
DDG	HFR	1237.47804	3
DDG	STR	1344.49555	4
MET	HFR	1211.64150	5
MET	STR	1350.49235	6

Least Squares Means for effect Calan.Treatment*SEX
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Final_Calf_BW

i/j	1	2	3	4	5	6
1		0.4073	0.9813	0.2665	0.7023	0.2691
2	0.4073		0.1507	1.0000	0.0196	0.9999
3	0.9813	0.1507		0.0869	0.9879	0.0913
4	0.2665	1.0000	0.0869		0.0070	1.0000
5	0.7023	0.0196	0.9879	0.0070		0.0094
6	0.2691	0.9999	0.0913	1.0000	0.0094	

Calan.Treatment	SEX	Final_Calf_BW LSMEAN	95% Confidence Limits	
CON	HFR	1265.707529	1213.707156	1317.707902
CON	STR	1341.334740	1282.436522	1400.232958
DDG	HFR	1237.478039	1177.025847	1297.930231
DDG	STR	1344.495553	1293.791878	1395.199229
MET	HFR	1211.641501	1158.339624	1264.943378
MET	STR	1350.492348	1292.107359	1408.877337

Figure 30: Tukey's HSD test and comparisons for ANCOVA model.