

# 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](https://github.com/RyanLalicker/Data-Analysis-2-STAT-325-825) 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 least 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.

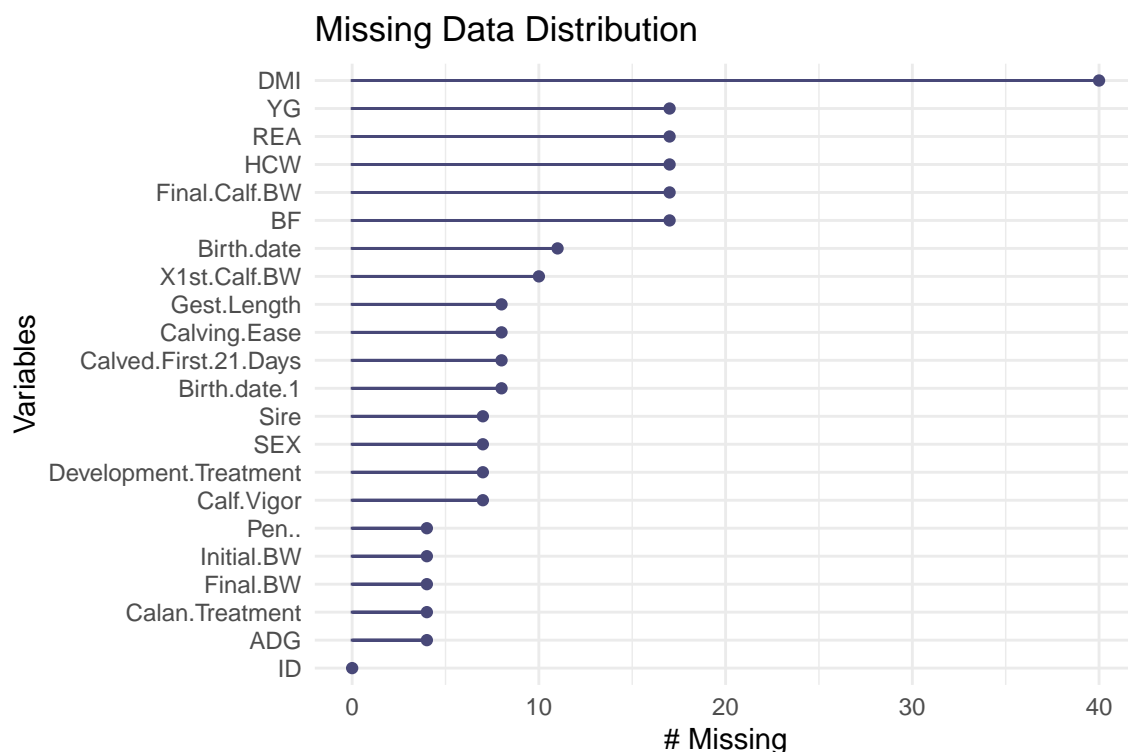


Figure 1: Chart counting the number of missing values for each variable within the data.

## 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.

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

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 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.

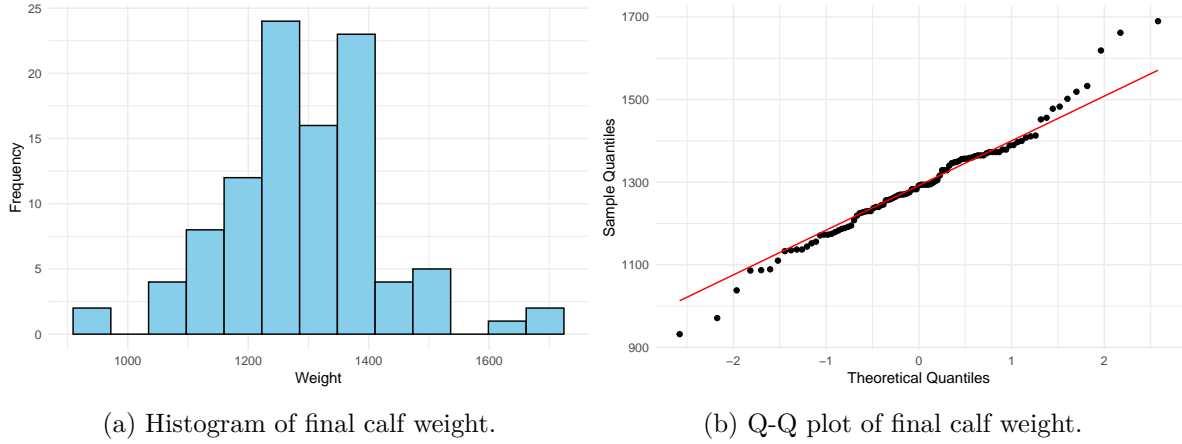


Figure 3: Plots used to check if the distribution of the final calf weight is normal.

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.

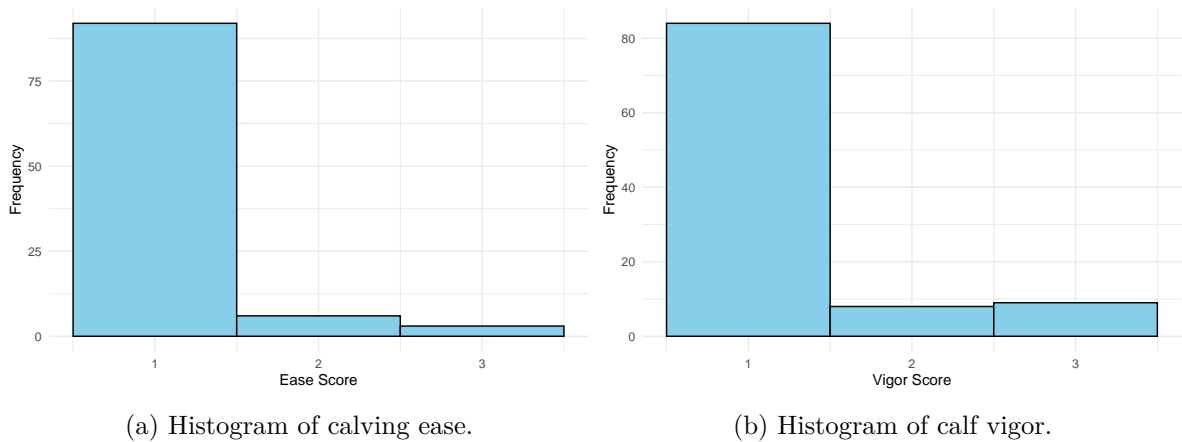


Figure 4: Histograms of scoring variables.

## 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.

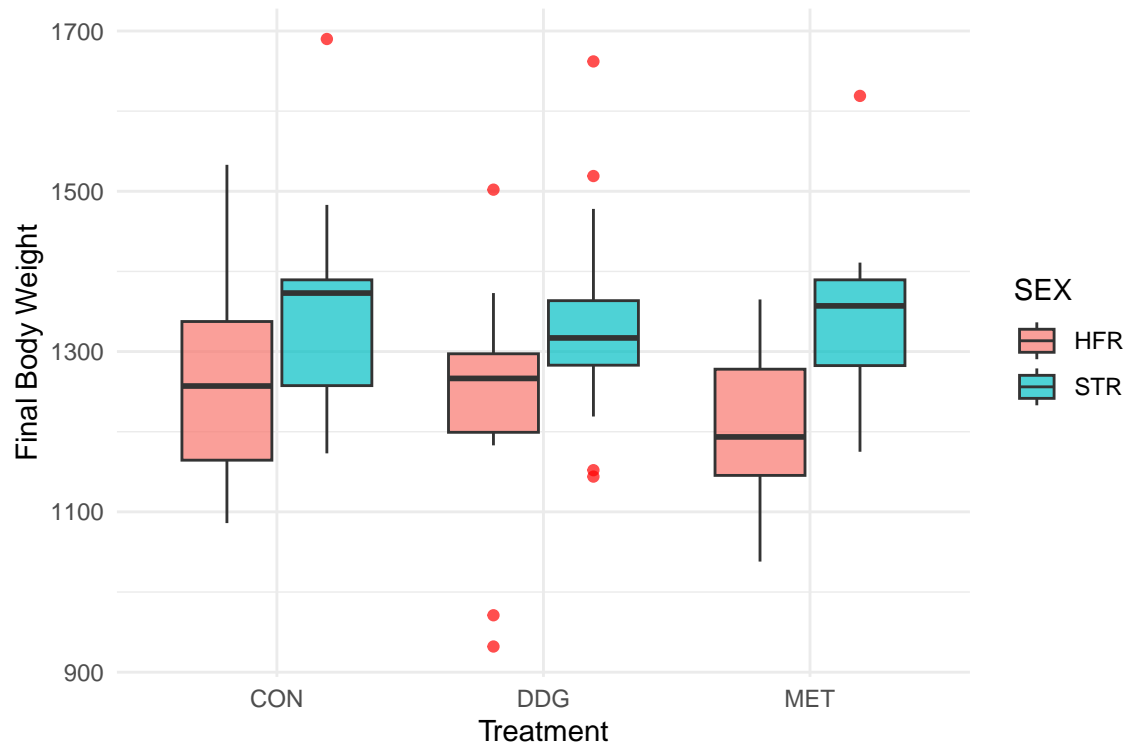


Figure 7: Final body weight by treatment and sex.

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.

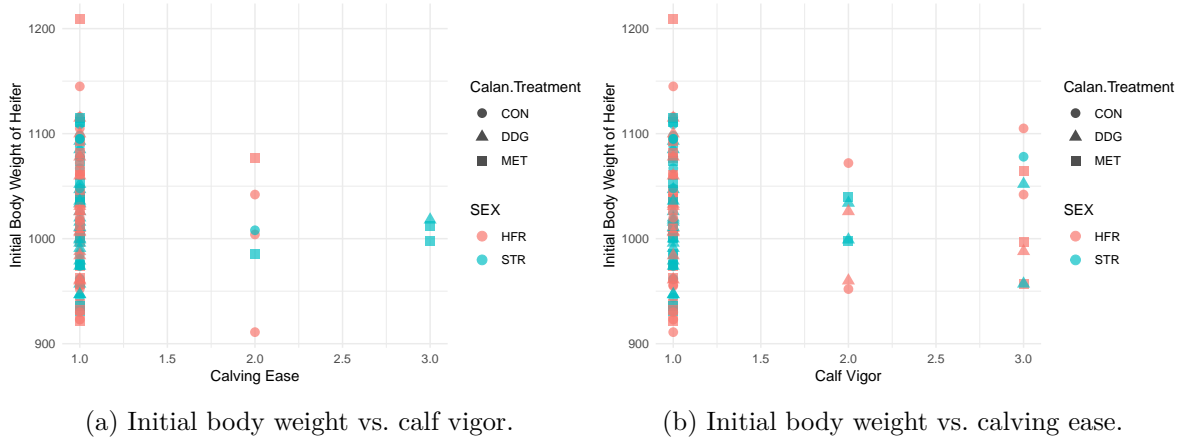


Figure 8: Scatterplot of heifer's initial body weight versus scoring variables, controlling for third trimester treatment and sex of the calf.

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

### 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  $l = 1, 2, \dots, 5$  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  $\alpha_l$  is the intercept of the model at the  $l$ th level,  $X_{trt}$  and  $X_{sex}$  represent the values for the third trimester treatment and the sex respectively, and the  $\beta_i$  are slope values corresponding to the  $i$ th 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:SEXSTR** represents the interaction effect between the DDG treatment and steers, which is highly significant (p-value < 0.001). The interaction between the MET treatment and steers is not significant though (p-value = 0.149). This means we can focus on the main effect for the MET treatment, but this is also not significant (p-value = 0.353).

	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 intercepts of the model. For instance the row 1|2 means for  $l \leq 1$  the model estimates  $\alpha_1 = 1.735$  and for  $l \leq 2$  we can say  $\alpha_2 = 2.939$ . Both of these intercepts 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 coefficient, 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 receive an AIC of 79.95121 as well, which seemed decent to us. (“Ordinal Logistic Regression | r Data Analysis Examples” (2011)).

In search of another model we considered a multinomial logistic regression. We have concerns with this approach however since this approach is designed to model nominal data, which is categorical variables without an order. (Frost (2021)). For that reason we chose to go in another direction.

## MAYBE BINARY INSTEAD Binomial for Calving Ease

**Need to check this** 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. That means the equation can be written just as it was in the ordinal logistic regression, with the exception of  $l$  which now only has the levels *low* and *high*.

The reason we can use the same function is that

we can express the function exactly as we did for the ordinal logistic regression, since binomial regression uses the same

Call:

```
glm(formula = Calving.Ease.bin ~ Calan.Treatment + SEX, family = binomial,
     data = data2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.21646	0.65044	-3.408	0.000655 ***
Calan.TreatmentDDG	-1.54608	1.15281	-1.341	0.179876
Calan.TreatmentMET	0.02839	0.75534	0.038	0.970016
SEXSTR	0.41913	0.71454	0.587	0.557485

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

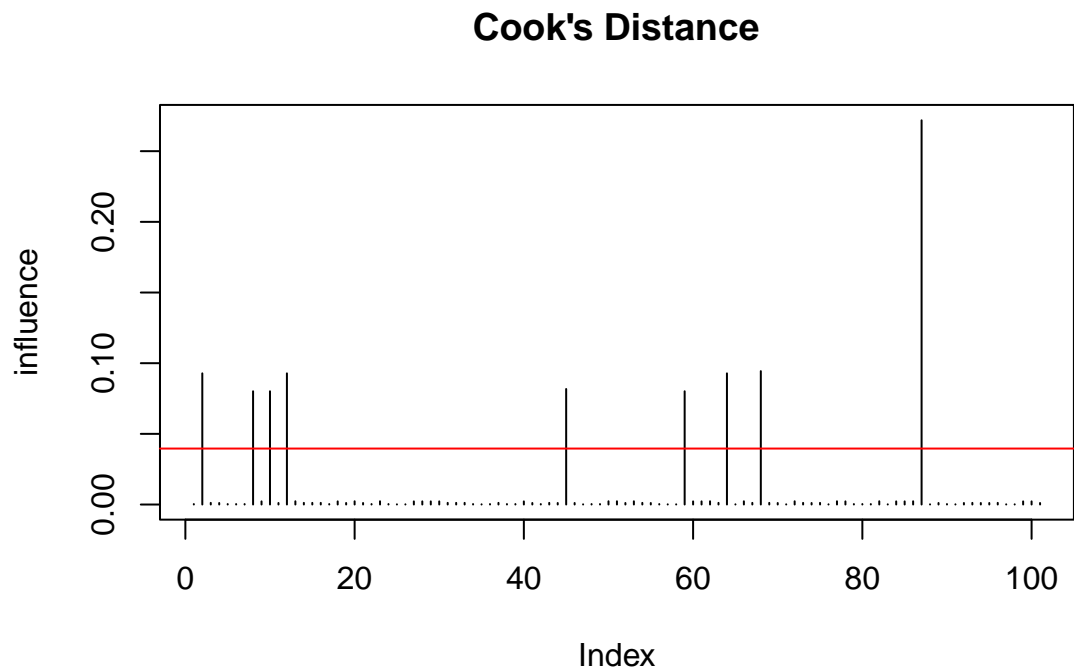
(Dispersion parameter for binomial family taken to be 1)

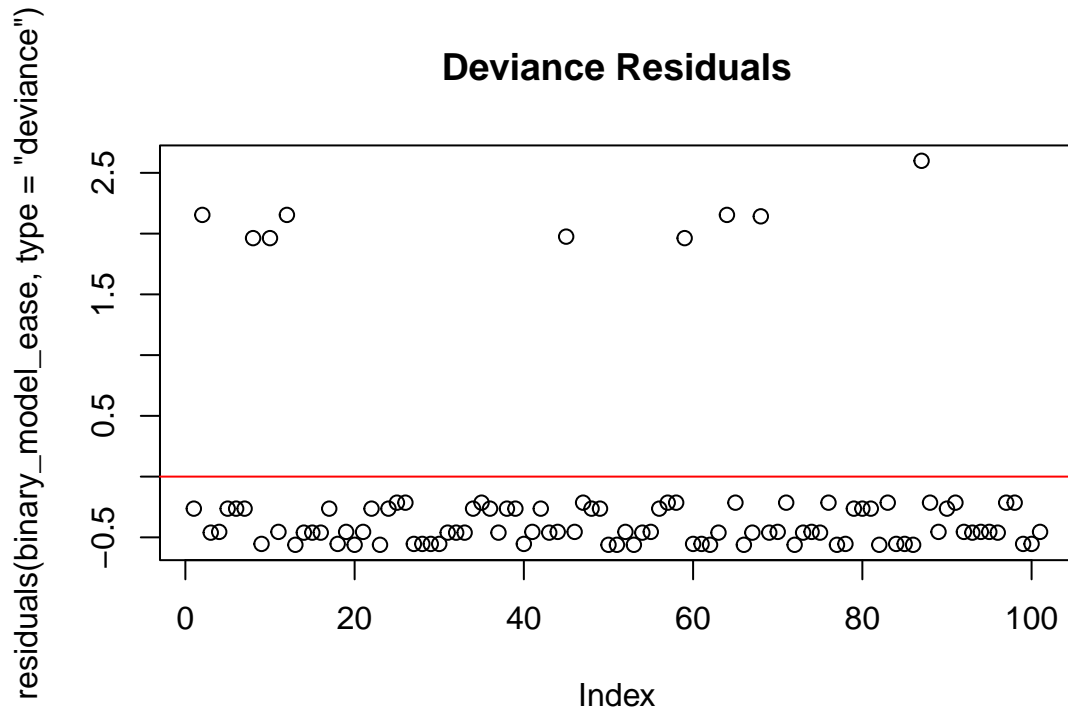
Null deviance: 60.695 on 100 degrees of freedom  
 Residual deviance: 57.682 on 97 degrees of freedom  
 AIC: 65.682

Number of Fisher Scoring iterations: 6

Checking Assumption for Binomial

	GVIF	Df	$GVIF^{1/(2 \cdot Df)}$
Calan.Treatment	1.010268	2	1.002557
SEX	1.010268	1	1.005121





## Models for Calf Vigor

### Ordinal Logistic Regression Model

**Interpretation** The ordinal logistic regression model for Calf.Vigor shows no statistically significant main effects for Calan.Treatment (DDG, MET) or SEX (STR vs. HFR), nor their interactions ( $p > 0.05$ ). The thresholds between vigor levels are significant ( $p = 0.017$  for 1 2,  $p = 0.001$  for 2 3), indicating that the ordinal levels are distinct, validating the use of this model. The model's residual deviance is 113.059 with an AIC of 127.059, suggesting reasonable fit but limited explanatory power for the predictors. Overall, the findings indicate that neither treatment nor sex strongly influences vigor scores, though the response levels remain ordinally structured. Further exploration of additional covariates may be needed to explain variability in Calf.Vigor

Call:

```
polr(formula = Calf.Vigor ~ Calan.Treatment * SEX, data = data2,
      Hess = TRUE)
```

Coefficients:

	Value	Std. Error	t value
Calan.TreatmentDDG	-0.02146	0.8519	-0.02519
Calan.TreatmentMET	-0.17876	0.8454	-0.21145
SEXSTR	-1.27076	1.1768	-1.07981
Calan.TreatmentDDG:SEXSTR	1.22910	1.4516	0.84673
Calan.TreatmentMET:SEXSTR	0.83354	1.5350	0.54301

Intercepts:

	Value	Std. Error	t value
1 2	1.3316	0.5592	2.3812
2 3	2.0682	0.6012	3.4399

Residual Deviance: 113.059

AIC: 127.059

	Value	Std. Error	t value	p-value
Calan.TreatmentDDG	-0.02146293	0.8519098	-0.02519389	0.9799003072
Calan.TreatmentMET	-0.17875927	0.8453950	-0.21145059	0.8325356827
SEXSTR	-1.27076073	1.1768363	-1.07981095	0.2802263761
Calan.TreatmentDDG:SEXSTR	1.22909789	1.4515849	0.84672823	0.3971466190
Calan.TreatmentMET:SEXSTR	0.83353681	1.5350294	0.54301033	0.5871226874
1 2	1.33164741	0.5592312	2.38121078	0.0172558351
2 3	2.06823750	0.6012463	3.43991700	0.0005818926

**Improvement of the model** This model is not improving also

Call:

```
polr(formula = Calf.Vigor ~ Calan.Treatment * SEX + Pen.. + Sire +
      Initial.BW, data = data2, Hess = TRUE)
```

Coefficients:

Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
-0.084807984	-0.440791870	-1.535799761
Pen..	Sire1616	Sire2213
0.029480559	0.563220216	0.194747258
Sire261	Sire4705	Sire4708
-16.184433315	-0.707327261	-0.482306675
Initial.BW	Calan.TreatmentDDG:SEXSTR	Calan.TreatmentMET:SEXSTR
0.003581125	1.299302274	1.205517814

Intercepts:

1|2      2|3  
4.836055 5.595344

Residual Deviance: 108.3906  
AIC: 136.3906

	Value	Std. Error	t value	p-value
Calan.TreatmentDDG	-0.084807984	7.663529e-01	-1.106644e-01	9.118825e-01
Calan.TreatmentMET	-0.440791870	7.481645e-01	-5.891643e-01	5.557510e-01
SEXSTR	-1.535799761	5.176354e-01	-2.966953e+00	3.007673e-03
Pen..	0.029480559	2.482094e-01	1.187729e-01	9.054553e-01
Sire1616	0.563220216	7.603037e-01	7.407833e-01	4.588249e-01
Sire2213	0.194747258	7.175354e-01	2.714114e-01	7.860747e-01
Sire261	-16.184433315	1.449370e-07	-1.116653e+08	0.000000e+00
Sire4705	-0.707327261	1.210301e+00	-5.844228e-01	5.589359e-01
Sire4708	-0.482306675	9.290051e-01	-5.191647e-01	6.036459e-01
Initial.BW	0.003581125	9.252225e-04	3.870556e+00	1.085876e-04
Calan.TreatmentDDG:SEXSTR	1.299302274	6.993823e-01	1.857786e+00	6.319948e-02
Calan.TreatmentMET:SEXSTR	1.205517814	7.198695e-01	1.674634e+00	9.400611e-02
1 2	4.836054937	1.226942e-01	3.941551e+01	0.000000e+00
2 3	5.595343903	2.843240e-01	1.967947e+01	3.233895e-86

**Comparison** Since they're not significant. No need to compare. From the VIF score, there's multicollinearity for Sex variable

SEX = HFR:

contrast	estimate	SE	df	z.ratio	p.value
CON - DDG	0.0215	0.852	Inf	0.025	0.9997
CON - MET	0.1788	0.845	Inf	0.211	0.9757
DDG - MET	0.1573	0.904	Inf	0.174	0.9834

SEX = STR:

contrast	estimate	SE	df	z.ratio	p.value
CON - DDG	-1.2076	1.175	Inf	-1.028	0.5594
CON - MET	-0.6548	1.281	Inf	-0.511	0.8660
DDG - MET	0.5529	0.937	Inf	0.590	0.8253

P value adjustment: tukey method for comparing a family of 3 estimates

## WE NEED TO CUT OR SUMMARIZE THIS - Multinomial Logistic Regression

The multinomial logistic regression model for Calf.Vigor with three levels (1,2,3) considers predictors Calan.Treatment (Con,DDG, MET), SEX (STR, HFR), and their interactions. None of the predictors or interaction terms are statistically significant, as all coefficients have large standard errors, indicating instability due to data sparsity. The residual deviance (104.3791) and AIC (128.3791) suggest a moderate fit, but high uncertainty in some estimates (e.g., for MET and SEXSTR) limits interpretability. The sparse combinations of response levels with predictors likely caused the inflated standard errors and poor precision. Addressing sparsity (e.g., collapsing categories or balancing the dataset) and simplifying the model (e.g., removing interactions) are essential next steps for improving the analysis

```
# weights:  21 (12 variable)
initial value 110.959841
iter  10 value 53.136150
iter  20 value 52.192910
iter  30 value 52.189566
final value 52.189559
converged
```

Call:

```
multinom(formula = Calf.Vigor ~ Calan.Treatment * SEX, data = data2)
```

Coefficients:

	(Intercept)	Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
2	-2.014703	0.3098633	-10.204618	-11.653191
3	-2.014924	-0.3829679	0.405398	-0.624152
	Calan.TreatmentDDG:SEXSTR	Calan.TreatmentMET:SEXSTR		
2	11.2786344	22.00063		
3	0.9425919	-12.94747		

Std. Errors:

	(Intercept)	Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
2	0.7527132	1.075883	116.2395818	248.246686
3	0.7527868	1.287470	0.9832157	1.279897
	Calan.TreatmentDDG:SEXSTR	Calan.TreatmentMET:SEXSTR		
2	248.249009	274.1132		
3	1.814259	548.9998		

Residual Deviance: 104.3791

AIC: 128.3791

Even without interaction this model is not fitting good

```
# weights: 15 (8 variable)
initial value 110.959841
iter 10 value 56.476935
final value 56.476752
converged
```

Call:

```
multinom(formula = Calf.Vigor ~ Calan.Treatment + SEX, data = data2)
```

Coefficients:

	(Intercept)	Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
2	-2.606433	0.7877341	0.03717386	-0.1537903
3	-1.988037	0.1836207	0.04510540	-0.7619077

Std. Errors:

	(Intercept)	Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
2	0.7975514	0.9138570	1.0347155	0.7521723
3	0.6487900	0.8704485	0.8628715	0.7470578

Residual Deviance: 112.9535

AIC: 128.9535

	df	AIC
vigor_model_1	7	127.0590
vigor_model_2	12	128.3791

## Checking Assumption

From the VIF score, there's multicollenearity for Sex variable

For Multinomial- The assumption of linearity in the log-odds for Final Calf Body Weight is not satisfied. This non-linearity suggests the need for transformations or more flexible modeling techniques.

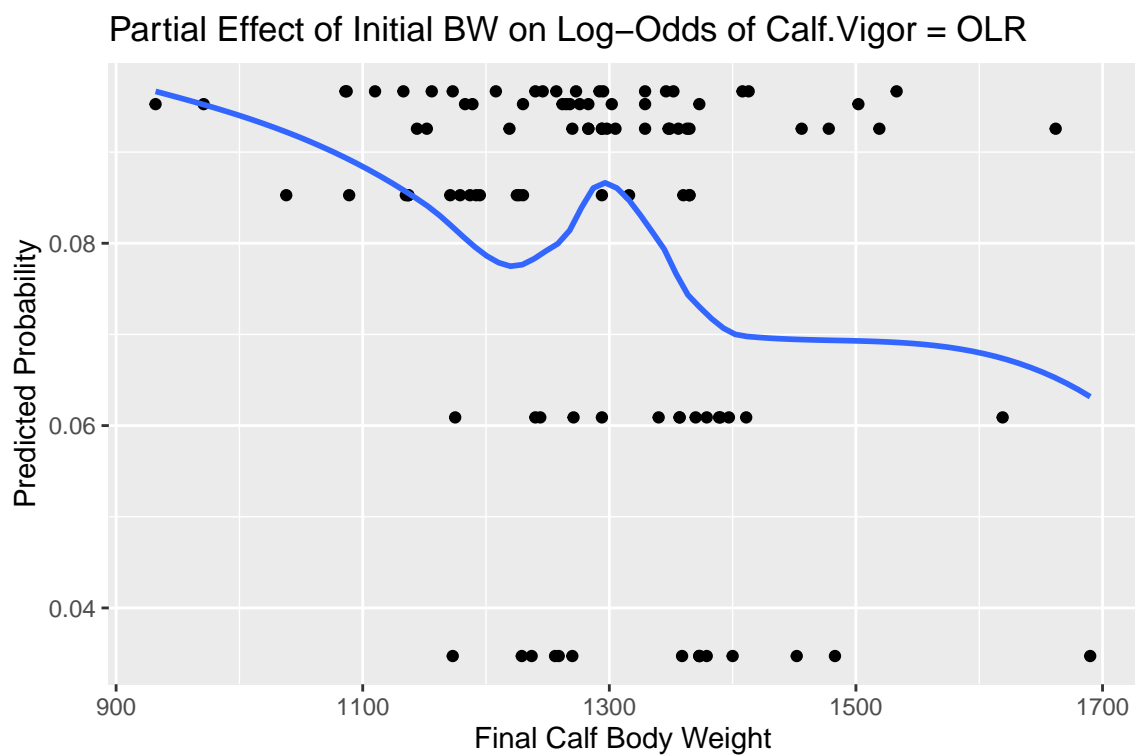
The Ordinal Logistic Regression (OLR) model shows a smoother relationship between Final Body Weight and log-odds of Calf.Vigor = 2, with stable predictions even in sparse regions. In contrast, the Multinomial Logistic Regression (MLR) model demonstrates greater variability, with sharp fluctuations influenced by sparse data at the extremes, capturing more detailed patterns. OLR handles data sparsity better and assumes proportional odds, making it simpler and more interpretable. MLR provides flexibility for datasets with complex structures but

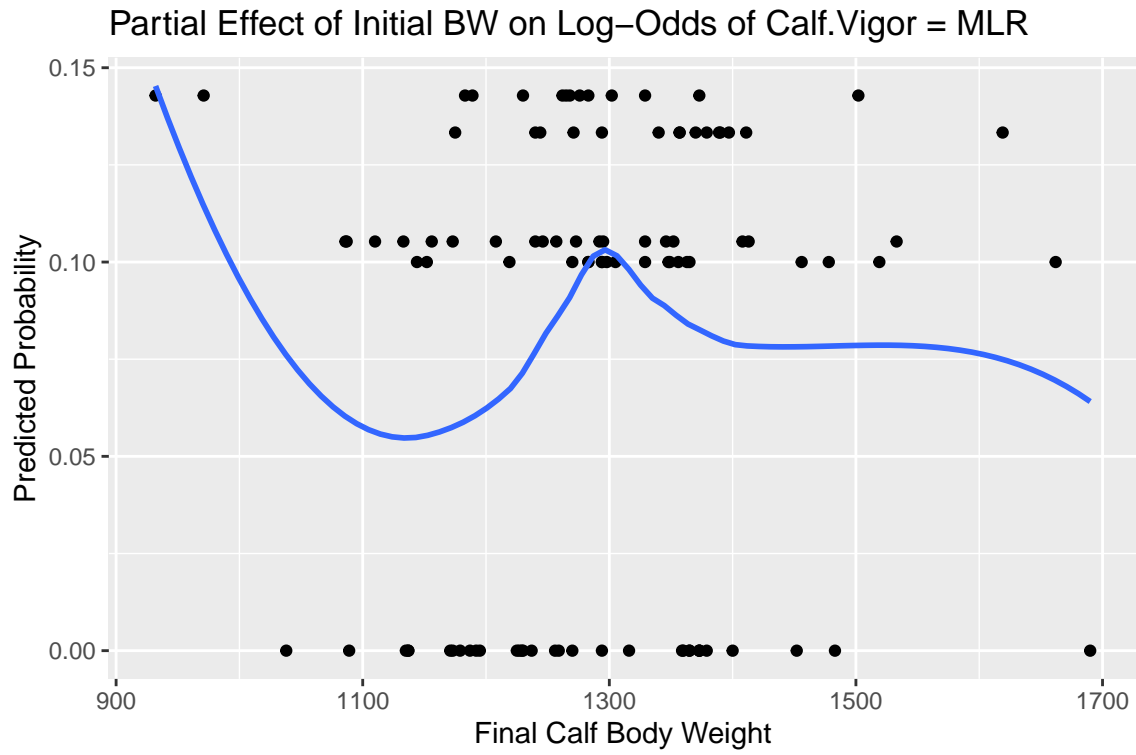


risks overfitting in regions with fewer observations. Overall, OLR is preferred for stability and interpretability, while MLR is more appropriate if the proportional odds assumption is violated or additional complexity is necessary.

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Calan.Treatment	2.982394	2	1.314139
SEX	4.744256	1	2.178131
Calan.Treatment:SEX	11.188273	2	1.828903

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Calan.Treatment	9.526533e+04	2	17.56846
SEX	2.177453e+05	1	466.63188
Calan.Treatment:SEX	2.598154e+09	2	225.76999





, , = HFR

	CON	DDG	MET
1	15	11	15
2	2	2	0
3	2	1	3

, , = STR

	CON	DDG	MET
1	14	16	13
2	0	2	2
3	1	2	0

**Binomial Model for Calf Vigor**

```
Call:
glm(formula = Calf.Vigor ~ Calan.Treatment + SEX, family = binomial,
     data = data2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.57031	0.52440	-2.994	0.00275 **
Calan.TreatmentDDG	0.48002	0.65209	0.736	0.46166
Calan.TreatmentMET	0.04144	0.68799	0.060	0.95197
SEXSTR	-0.46790	0.54722	-0.855	0.39252

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

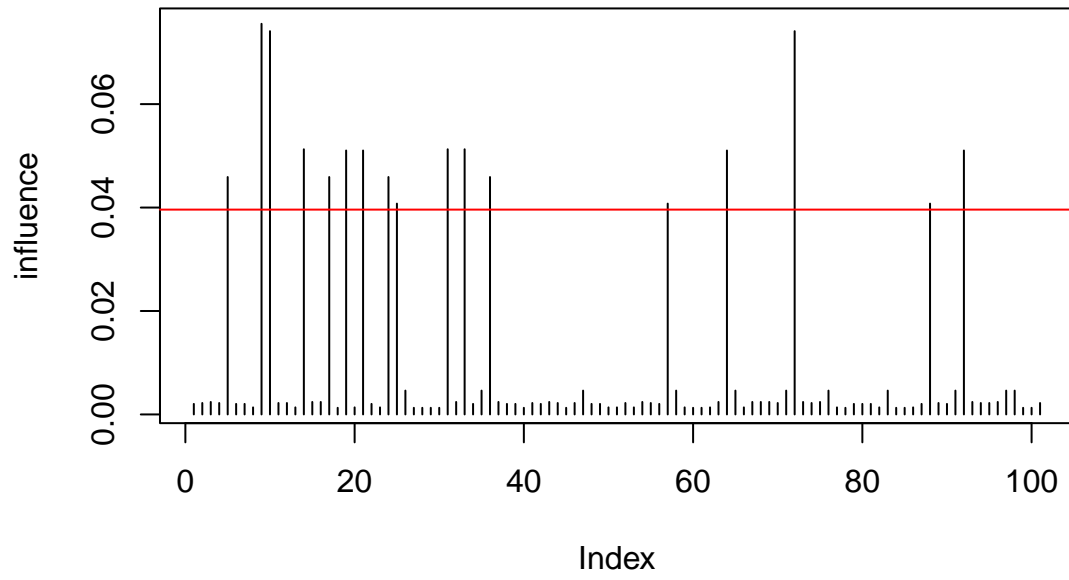
Null deviance: 91.548 on 100 degrees of freedom  
 Residual deviance: 90.298 on 97 degrees of freedom  
 AIC: 98.298

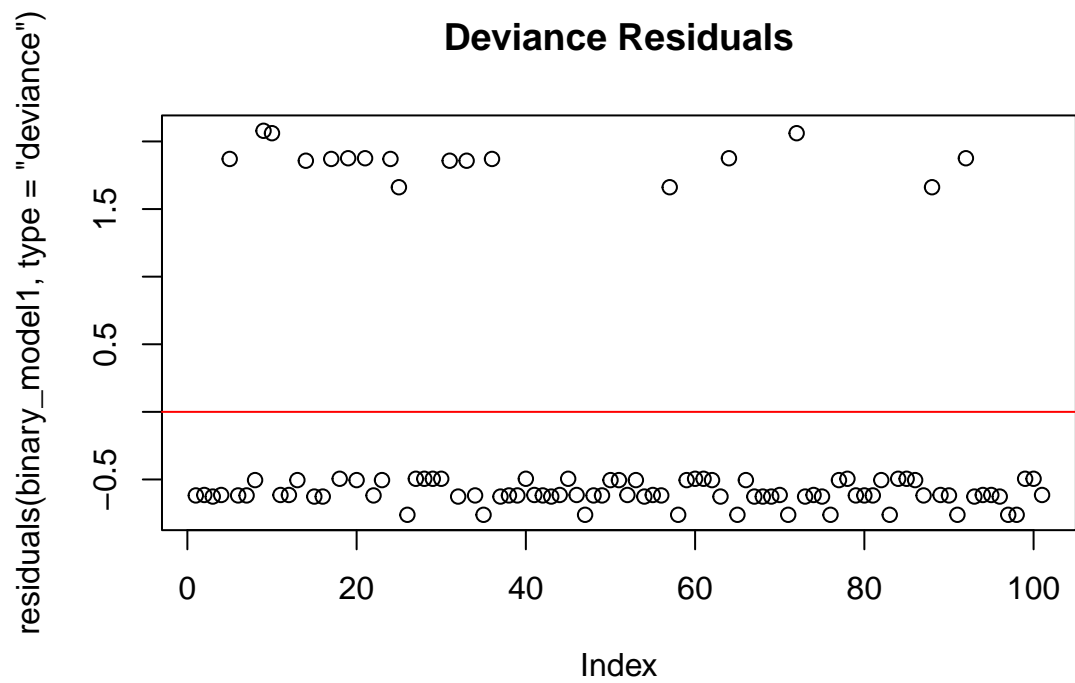
Number of Fisher Scoring iterations: 4

### Checking Assumption for Calf Vigor

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Calan.Treatment	1.023736	2	1.005882
SEX	1.023736	1	1.011798

## Cook's Distance





## Models for Final Calf Weight

### Mixed Model

will have to add SAS code.

SAS output

Estimated G matrix is not positive definite.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	Pen..	0
Intercept	Sire	2371.13
Residual		12485

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

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Calan.Treatment	2	87	0.41	0.6661
SEX	1	87	25.16	<.0001
Calan.Treatment*SEX	2	87	0.23	0.7919

Least Squares Means							
Effect	Calan.Treatment	SEX	Estimate	Standard Error	DF	t Value	Pr >  t
Calan.Treatment*SEX	CON	HFR	1270.83	33.1165	87	38.37	<.0001
Calan.Treatment*SEX	CON	STR	1363.88	35.4115	87	38.52	<.0001
Calan.Treatment*SEX	DDG	HFR	1236.86	36.2868	87	34.09	<.0001
Calan.Treatment*SEX	DDG	STR	1353.65	32.6267	87	41.49	<.0001
Calan.Treatment*SEX	MET	HFR	1230.74	33.5410	87	36.69	<.0001
Calan.Treatment*SEX	MET	STR	1361.54	36.0112	87	37.81	<.0001

Figure 10: Fig-1

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	-93.0475	38.9263	87	-2.39	0.0190	Tukey-Kramer	0.1711
Calan.Treatment* SEX	CON	HFR	DDG	HFR	33.9686	40.4263	87	0.84	0.4031	Tukey-Kramer	0.9591
Calan.Treatment* SEX	CON	HFR	DDG	STR	-82.8190	36.4150	87	-2.27	0.0254	Tukey-Kramer	0.2158
Calan.Treatment* SEX	CON	HFR	MET	HFR	40.0895	37.2867	87	1.08	0.2853	Tukey-Kramer	0.8899
Calan.Treatment* SEX	CON	HFR	MET	STR	-90.7108	38.7582	87	-2.34	0.0215	Tukey-Kramer	0.1894
Calan.Treatment* SEX	CON	STR	DDG	HFR	127.02	41.9553	87	3.03	0.0032	Tukey-Kramer	0.0370
Calan.Treatment* SEX	CON	STR	DDG	STR	10.2285	38.2804	87	0.27	0.7899	Tukey-Kramer	0.9998
Calan.Treatment* SEX	CON	STR	MET	HFR	133.14	39.2923	87	3.39	0.0011	Tukey-Kramer	0.0131
Calan.Treatment* SEX	CON	STR	MET	STR	2.3368	41.1047	87	0.06	0.9548	Tukey-Kramer	1.0000
Calan.Treatment* SEX	DDG	HFR	DDG	STR	-116.79	39.3933	87	-2.96	0.0039	Tukey-Kramer	0.0438
Calan.Treatment* SEX	DDG	HFR	MET	HFR	6.1209	40.3192	87	0.15	0.8797	Tukey-Kramer	1.0000
Calan.Treatment* SEX	DDG	HFR	MET	STR	-124.68	42.5174	87	-2.93	0.0043	Tukey-Kramer	0.0477
Calan.Treatment* SEX	DDG	STR	MET	HFR	122.91	36.4355	87	3.37	0.0011	Tukey-Kramer	0.0137
Calan.Treatment* SEX	DDG	STR	MET	STR	-7.8917	38.5696	87	-0.20	0.8384	Tukey-Kramer	0.9999
Calan.Treatment* SEX	MET	HFR	MET	STR	-130.80	39.6148	87	-3.30	0.0014	Tukey-Kramer	0.0170

Figure 11: Fig-2

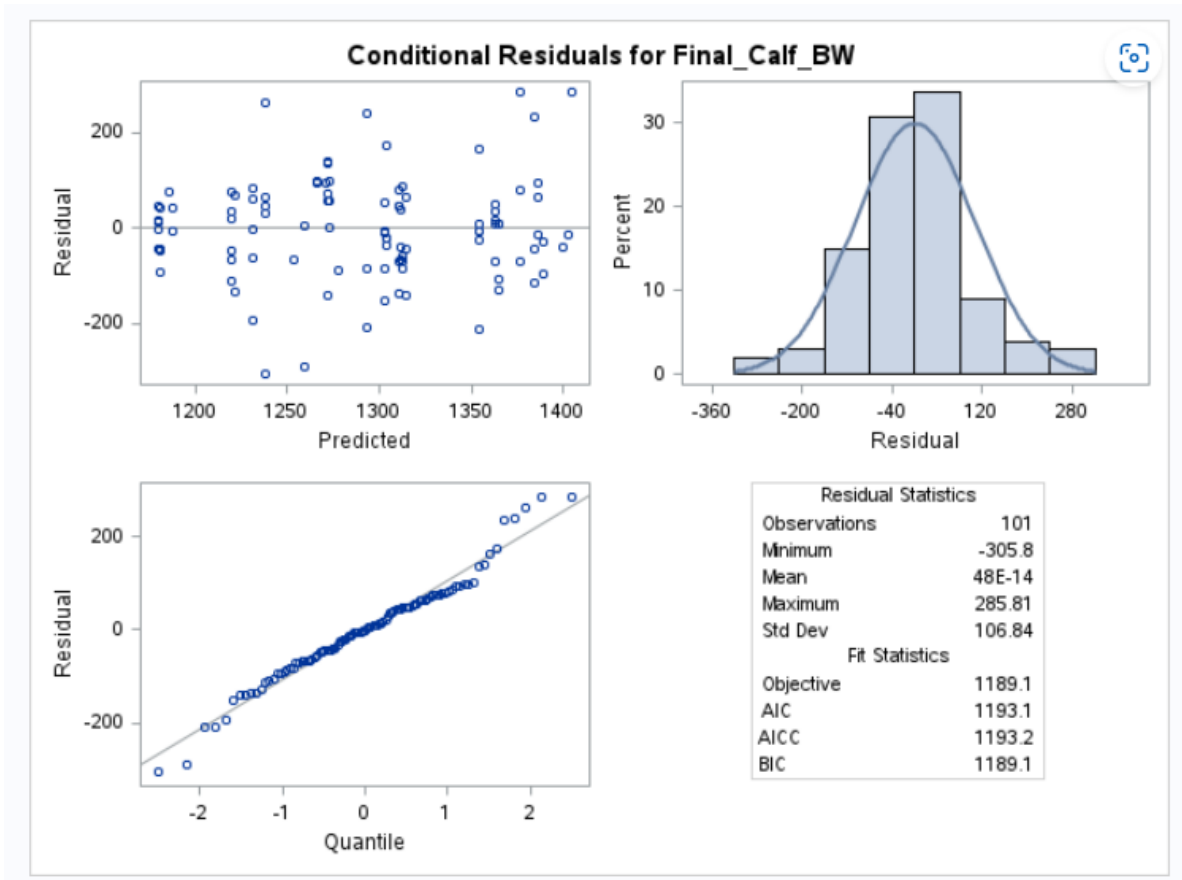


Figure 12: Fig-1

**Linear Mixed Model for Final Calf Body Weight** The linear mixed model shows that Final Calf Body Weight is significantly higher for Steers (STR) compared to Heifers (HFR) ( $p < 0.05$ ), but there are no significant differences between treatments (CON, DDG, MET) or their interactions with sex. The random effect for Sire contributes variance (237423742374), while Pen has negligible variance (000), indicating it does not affect body weight. Tukey HSD and pairwise comparisons confirm no significant differences between treatments within or across sexes ( $p > 0.05$ ). The high residual variance (124841248412484) suggests unaccounted variability, indicating that additional predictors might improve the model. Overall, treatment effects are not significant, but sex remains an important factor influencing final body weight.



## ANCOVA Model for Final Calf Weight

### 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

SAS Output

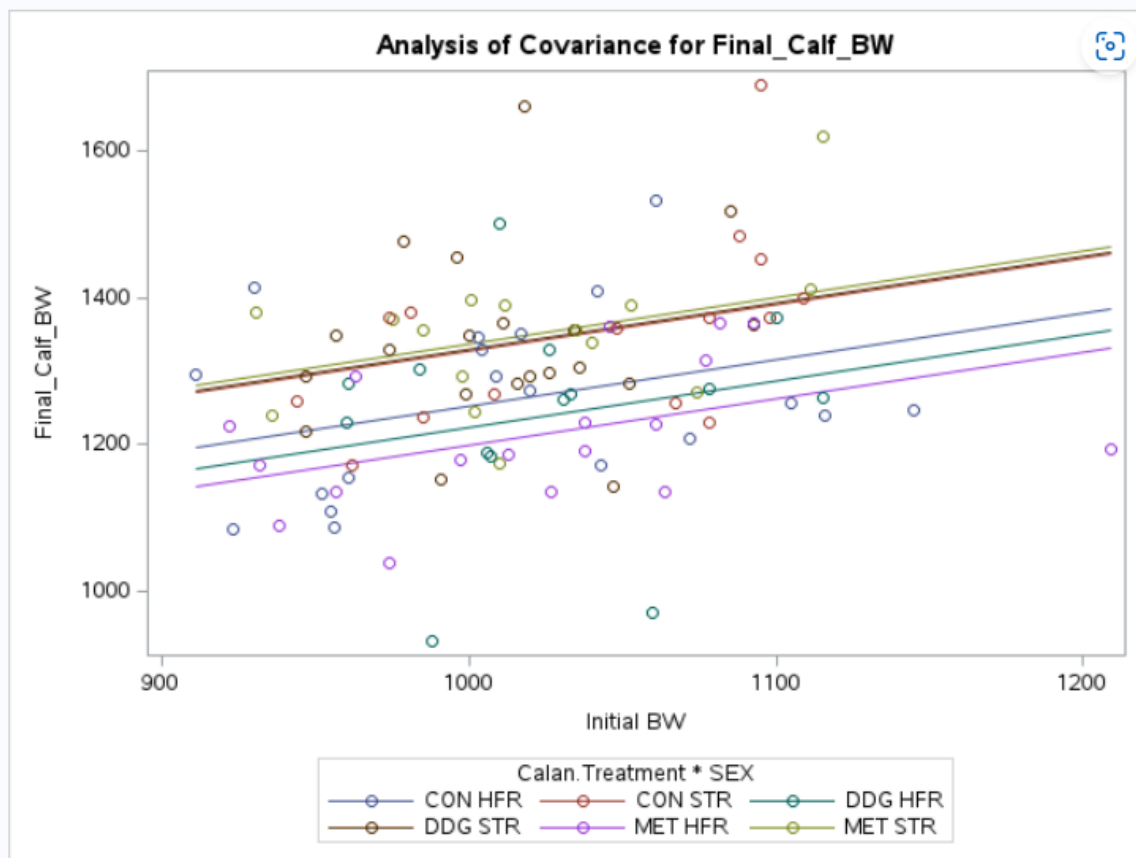


Figure 13: Fig-1

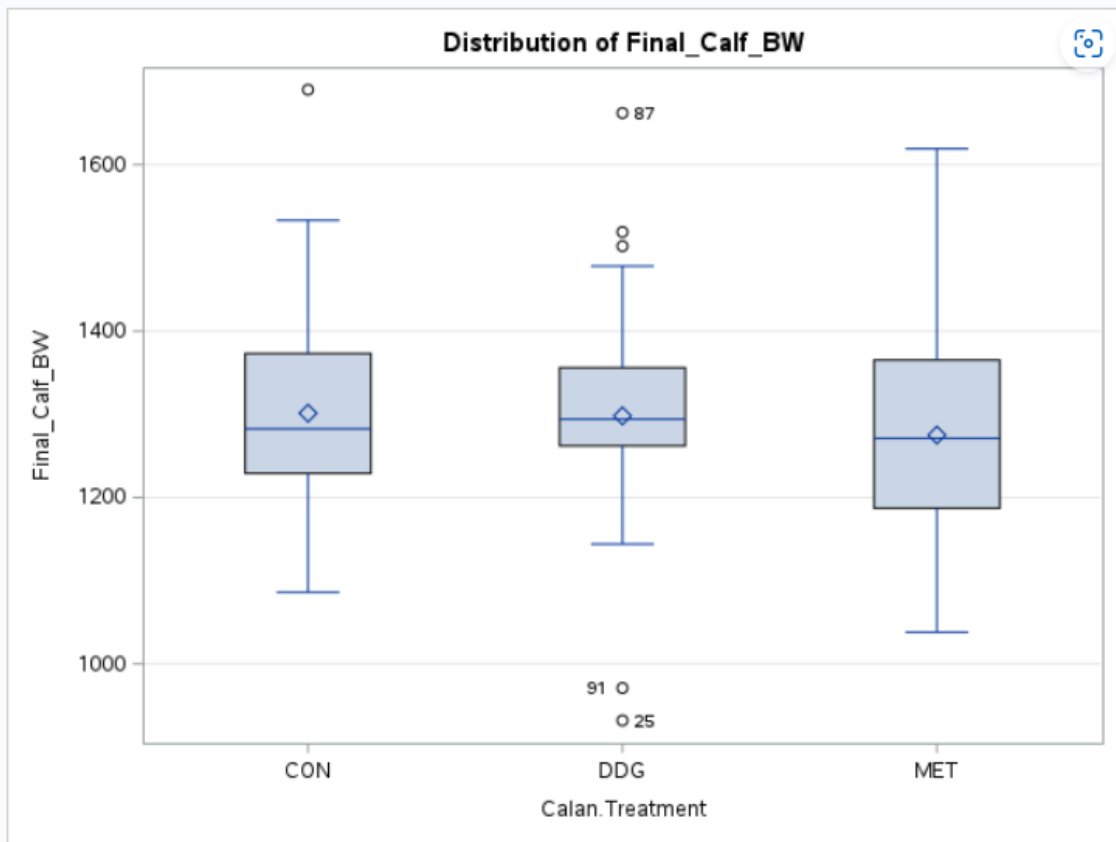


Figure 14: Fig-1

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 15: Fig-1

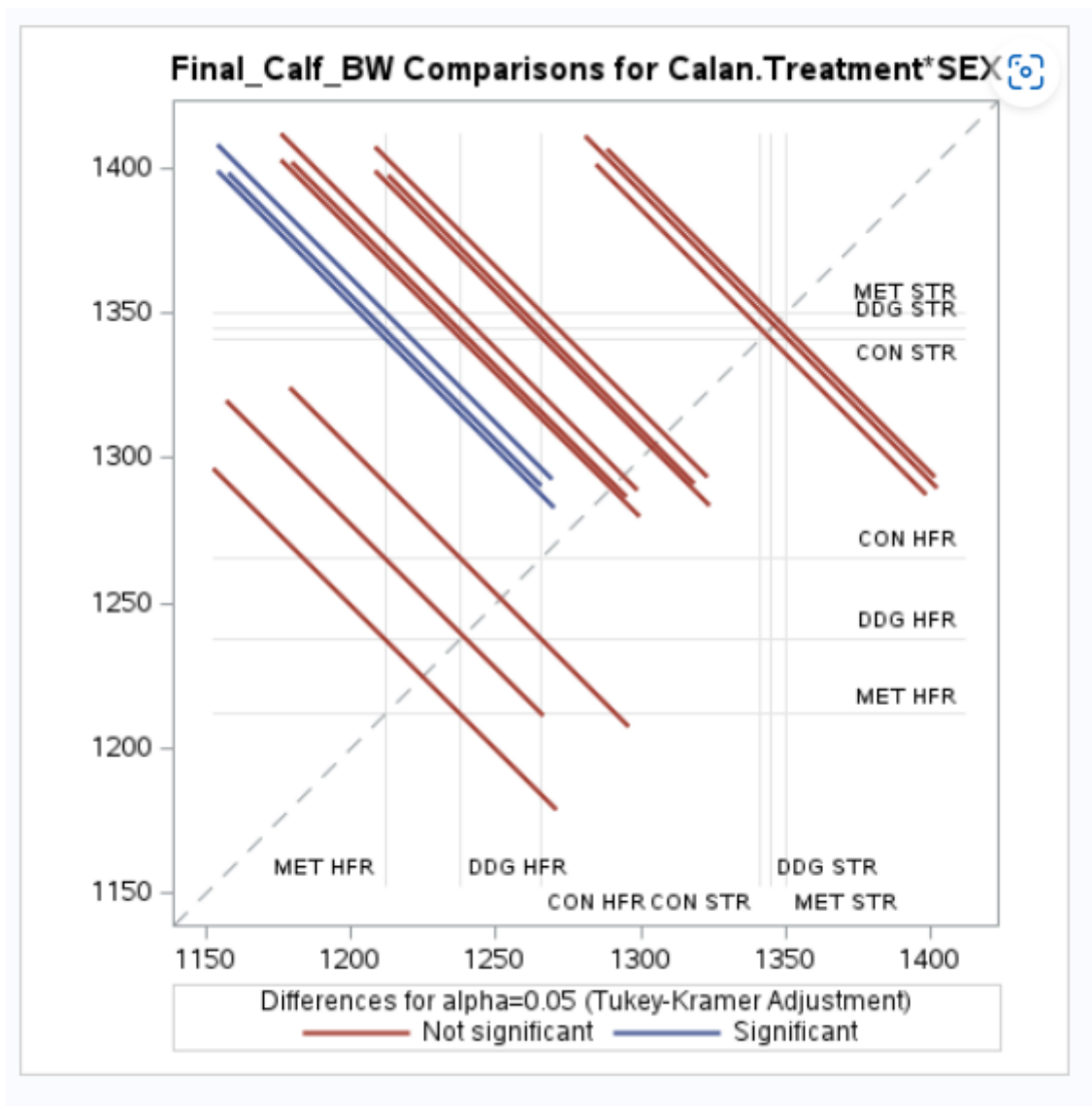


Figure 16: Fig-1

## Checking Assumptions

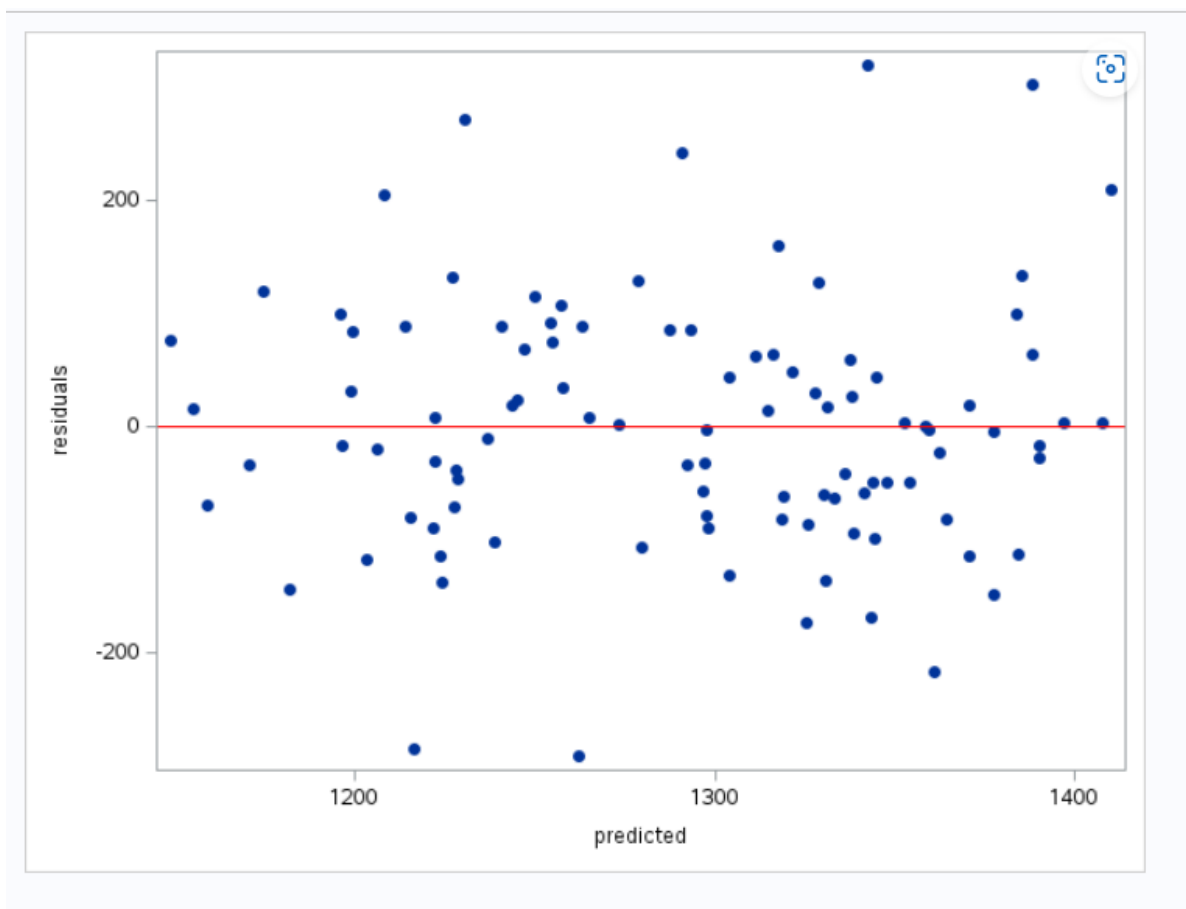


Figure 17: Fig-1

**The UNIVARIATE Procedure**  
**Variable: residuals**

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

Basic Statistical Measures			
Location		Variability	
<b>Mean</b>	0.00000	<b>Std Deviation</b>	110.39989
<b>Median</b>	-2.30039	<b>Variance</b>	12188
<b>Mode</b>	.	<b>Range</b>	610.59481
		<b>Interquartile Range</b>	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 18: Fig-1

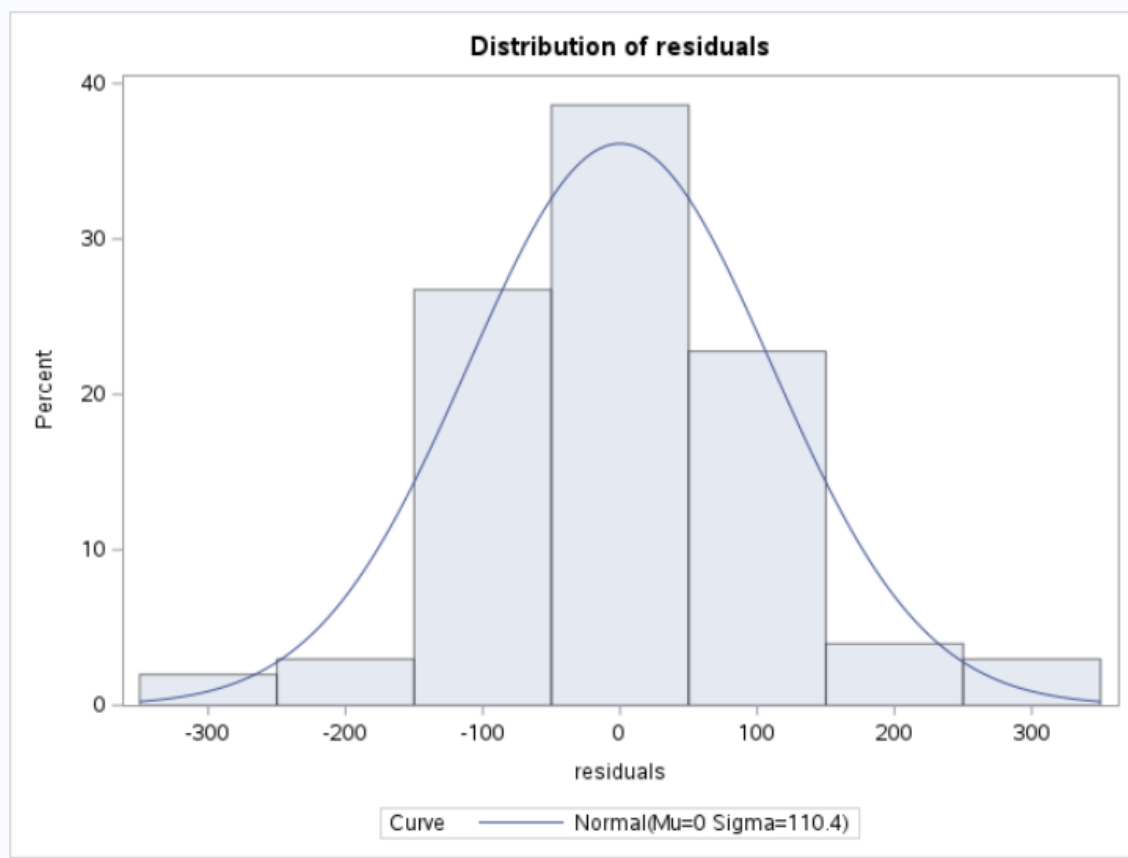


Figure 19: Fig-1



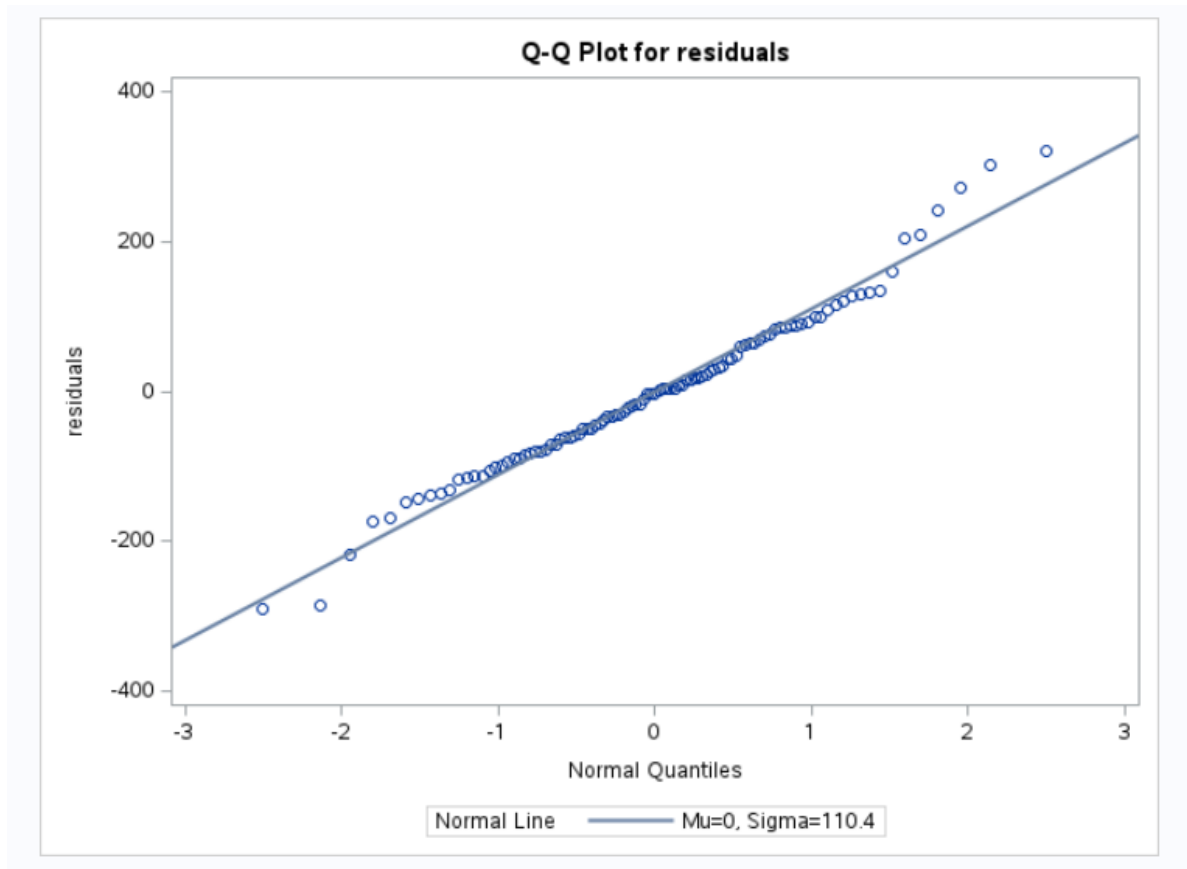


Figure 20: Fig-1

**If needed (for using later)**

**Model 1: LM for Final Calf Weight**

Comparison

Post Hoc Test

## 2. OLR Model

## 3. Binary Logistic Regression with GLM

### WE NEED TO CUT OR SUMMARIZE THIS - Multinomial Logistic Regression for Calf Calving Ease

The multinomial logistic regression model for Calving.Ease successfully converged, with a residual deviance of 58.42 and an AIC of 82.42, indicating moderate model fit. Extreme coefficients (e.g.,  $-89.53$  for Calan.TreatmentDDG) and large standard errors, especially for level 3, suggest data sparsity and instability in the estimates. Warnings for negative variance estimates point to multicollinearity or insufficient data for reliable interaction and treatment effect estimation. Post-hoc Tukey comparisons showed no significant differences between treatments ( $p=1.000$ ), and missing standard errors (NaN) highlight issues with data variability. Overall, the model's results are unreliable due to data limitations, necessitating simplified modeling or alternative approaches.

```
# weights:  21 (12 variable)
initial value 110.959841
iter  10 value 29.432881
iter  20 value 29.222854
iter  30 value 29.215412
iter  40 value 29.209442
iter  50 value 29.208986
final value 29.208984
converged
```

Call:

```
multinom(formula = Calving.Ease ~ Calan.Treatment * SEX, data = data2)
```

Coefficients:

	(Intercept)	Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
2	-1.673984	-89.53191	-1.159207	-0.9650507
3	-21.595074	4.64936	4.446292	1.2808117
	Calan.TreatmentDDG:SEXSTR	Calan.TreatmentMET:SEXSTR		
2	-6.615244	1.313325		
3	12.720477	14.076217		

Std. Errors:

	(Intercept)	Calan.TreatmentDDG	Calan.TreatmentMET	SEXSTR
2	0.6291544	6.210743e-07	1.206083	1.211298
3	604.7136553	9.531250e+02	955.622215	603.218409

	Calan.TreatmentDDG:SEXSTR	Calan.TreatmentMET:SEXSTR
2	7.567627e-08	1.899839
3	9.535708e+02	956.119462

Residual Deviance: 58.41797

AIC: 82.41797

SEX = HFR:

contrast	estimate	SE	df	t.ratio	p.value
CON - DDG	-9.39e-18	1.90e-10	12	0.000	1.0000
CON - MET	-4.49e-17	NaN	12	NaN	NaN
DDG - MET	-3.55e-17	NaN	12	NaN	NaN

SEX = STR:

contrast	estimate	SE	df	t.ratio	p.value
CON - DDG	-4.84e-17	3.09e-10	12	0.000	1.0000
CON - MET	-1.72e-17	NaN	12	NaN	NaN
DDG - MET	3.12e-17	NaN	12	NaN	NaN

Results are averaged over the levels of: Calving.Ease

P value adjustment: tukey method for varying family sizes

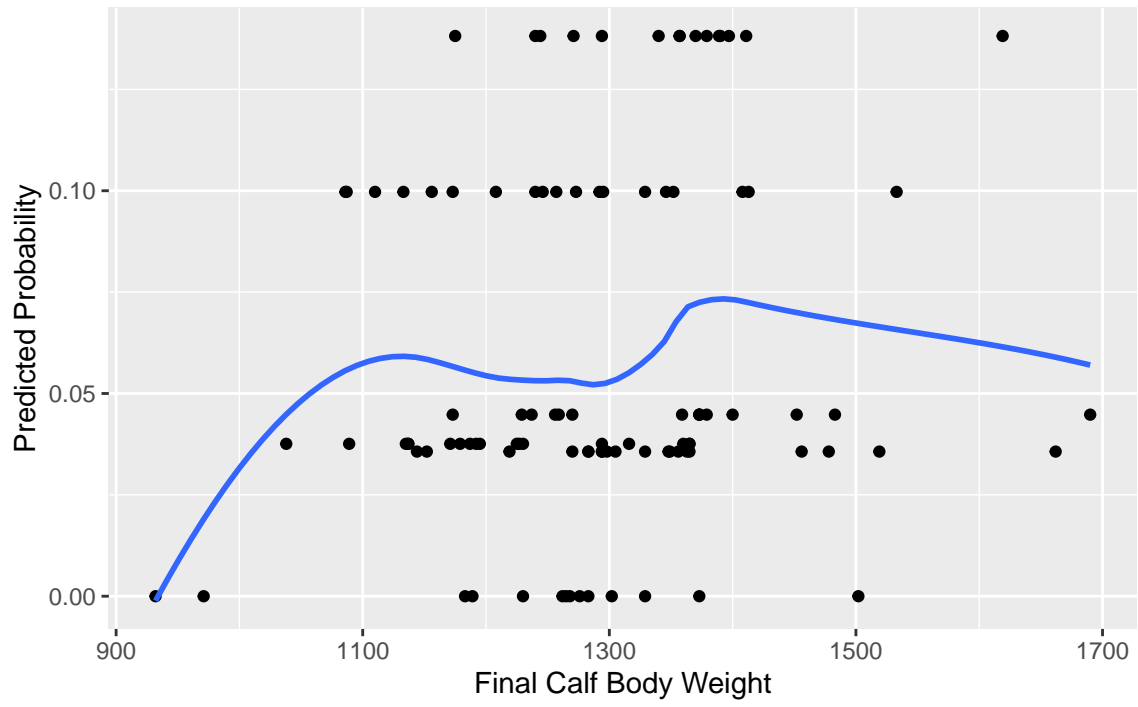
## Checking Assumption

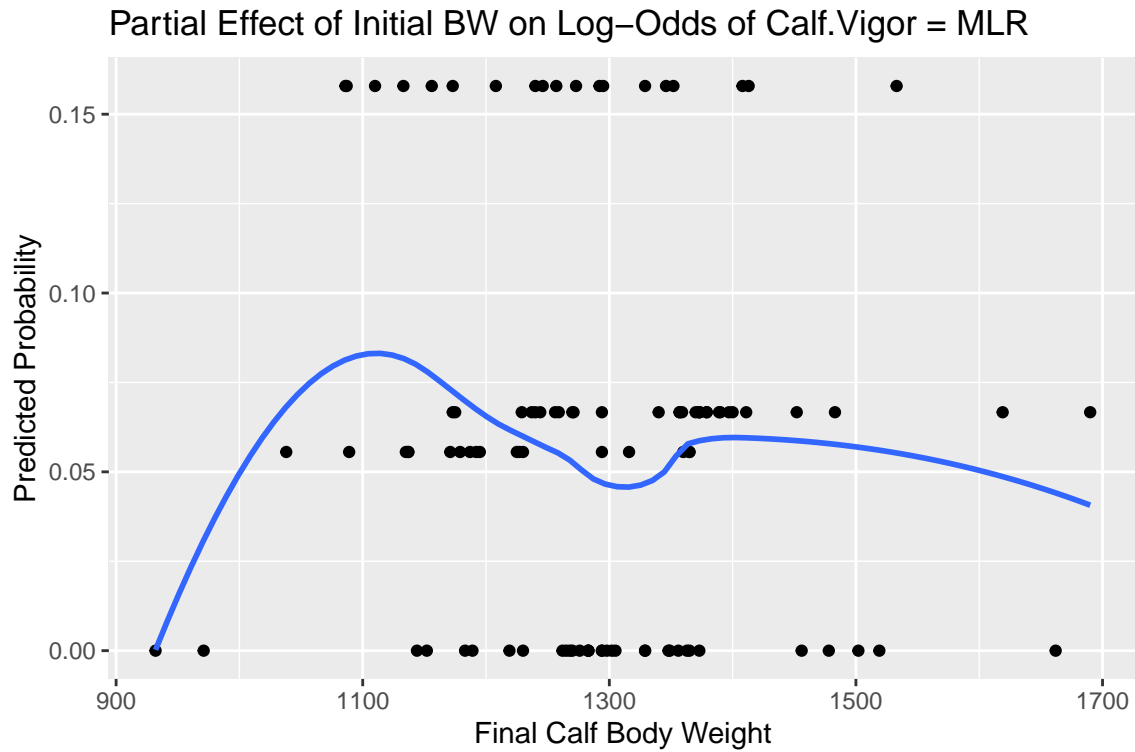
From VIF, Holding the assumptions for both model

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Calan.Treatment	7.052272e+15	2	9163.94049
SEX	1.282995e+02	1	11.32694
Calan.Treatment:SEX	1.014789e+16	2	10036.76805

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Calan.Treatment	-3.000797e+15	2	NaN
SEX	2.733875e+00	1	1.653443
Calan.Treatment:SEX	3.357161e+16	2	13536.083846

Partial Effect of Initial BW on Log-Odds of Calf.Vigor = OLR





## Equation Format

### Model 1

$$y_{ijklmn} = ENTER - MODEL - HERE$$

where  $y_{ijklm}$  represents the *dependent variable*, ...

! [Picture of SAS Output] (filename.png){width="3in"}

## Conclusion

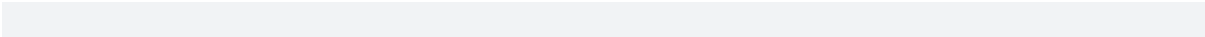
## Recomendation

## 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/>.
- 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>.
- 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>.
- 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

**Appendix B - SAS Code**





## Appendix C - Additional SAS Output

