Data Analysis 3

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

November 19, 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 entries, 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.

Exploring the Data

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 data set.

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 variable names line up with these descriptions.

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

UPDATE THIS AFTER SEEING WHAT VARIABLES ARE NEEDED FOR THE MODEL

The data contains some missing values. In total 53 rows in the data set 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

To clean up the data set let's first focus on some of the quantitative variables. For these we can impute the missing values with the median or mean. We will use the median due to the its greater resistance to potential outliers. This reduces the number of rows with at least one missing variable to 15. Much of the improvement comes from `DMI` being a quantitative variable, but to reiterate what was discussed earlier, this variable will not be used in any models.

The categorical variables are not as simple. 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 used the mode as Memon, Wamala, and Kabano (2023) suggests, this would make 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. For this reason, rows with missing values for the categorical variables used in the model will be removed. This results in the models using 113 entries of the original 120, which is something we are comfortable with.

RETURN AND VARIFY NUMBERS BEFORE SUBMITTING.

Summary Statistics

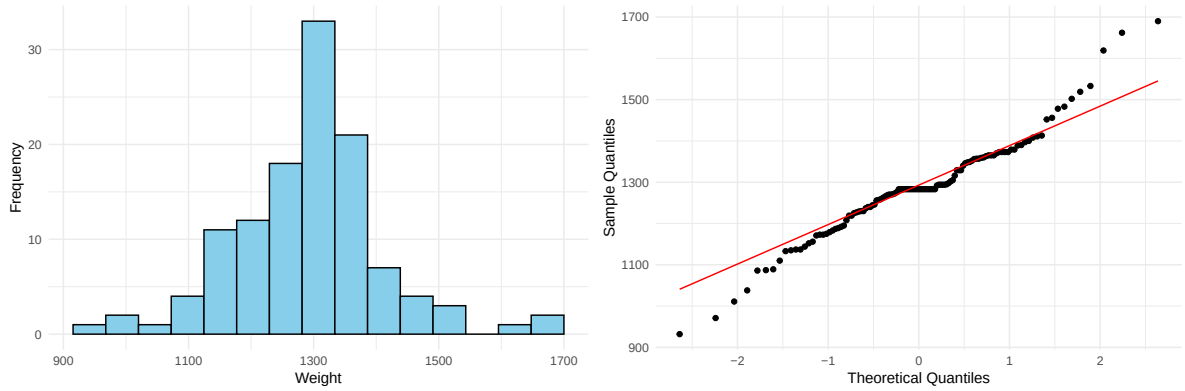
Let's take a closer look at what the three dependent variables the client is interested in. Figure 2 shows several summary statistics for each. The calving ease and calf vigor are each scores given. Looking at the minimum and maximum values of each it would seem they are scored in a three-point and five-point system respectively, both only using integers. We can also see from the median and 75th percentiles that both seem very skewed towards the low end of the scale. While the imputing done previously may be exaggerating the curve, Figure 1 shows less than ten values were imputed. This indicates to us the skew was already present before cleaning the data.

Variable	Mean	Median	SD	25th Percentile	75th Percentile	Min	Max
Calving.Ease	1.132743	1	0.4331039	1	1	1	3
Calf.Vigor	1.362832	1	0.8244256	1	1	1	5
Final.Calf.BW	1290.106195	1283	122.0959344	1227	1360	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 had some concerns on top of being counting variables on a scale, 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.

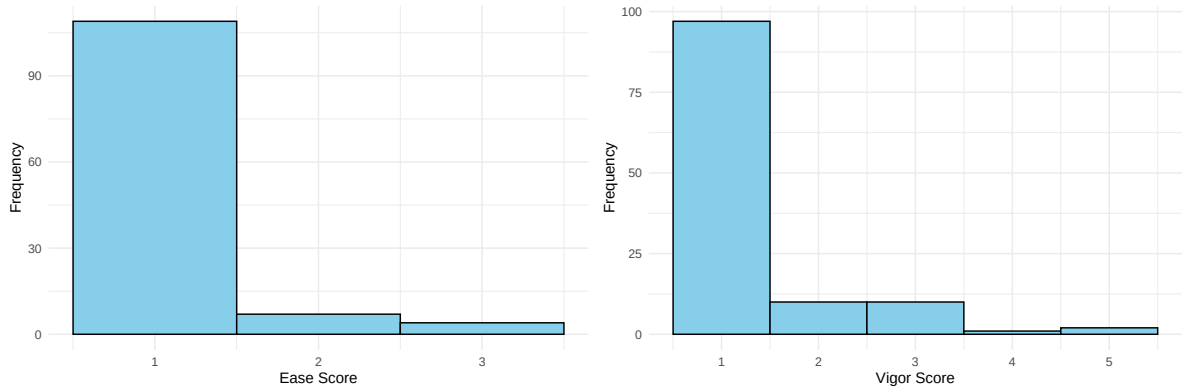


(a) Histogram of final calf weight.

(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.

	HFR	STR
CON	20	17
DDG	15	23
MET	19	19

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

Now let's consider how these variables affect the final body weight. The boxplot shown in Figure 6 indicates the highest median weights come from the MET treatment, followed closely by DDG. Additionally, the steers are heavier on average than the heifers. This indicates both likely have a use in predicting the final weight of a calf. The variance differs by both treatment and sex, but the only outlier is a heifer calf with the MET treatment.

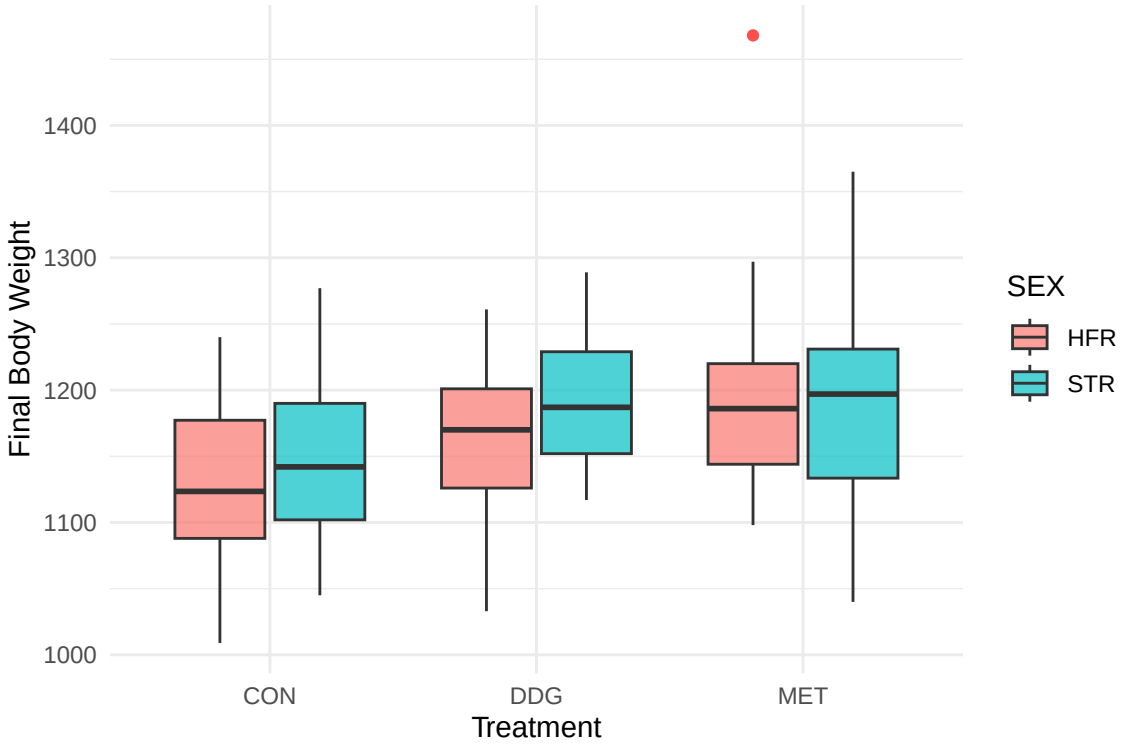


Figure 6: Final body weight by treatment and sex.

Models for Calving Ease

Ordinal Logistic Regression Model

```
Call:
polr(formula = Calving.Ease ~ Calan.Treatment * SEX, data = data2,
    Hess = TRUE)

Coefficients:
                            Value Std. Error  t value
Calan.TreatmentDDG        -16.7932     0.6341 -26.4836
Calan.TreatmentMET         -1.1127     1.2003  -0.9270
SEXSTR                     -0.2646     0.9739  -0.2717
Calan.TreatmentDDG:SEXSTR  15.7936     0.6341  24.9073
Calan.TreatmentMET:SEXSTR   1.9542     1.5240   1.2822

Intercepts:
    Value    Std. Error t value
1|2   1.7963   0.6222     2.8870
2|3   2.9177   0.7396     3.9449

Residual Deviance: 79.01625
AIC: 93.01625
```

	Value	Std. Error	t value	p-value
Calan.TreatmentDDG	-16.7931566	0.6340969	-26.4835793	1.498375e-154
Calan.TreatmentMET	-1.1127063	1.2002826	-0.9270369	3.539074e-01
SEXSTR	-0.2645934	0.9738552	-0.2716968	7.858551e-01
Calan.TreatmentDDG:SEXSTR	15.7936403	0.6340968	24.9073017	6.201312e-137
Calan.TreatmentMET:SEXSTR	1.9541929	1.5240486	1.2822379	1.997592e-01
1|2	1.7962564	0.6221810	2.8870320	3.888946e-03
2|3	2.9177029	0.7396133	3.9449031	7.983226e-05

Multinomial Logistic Regression

```
# weights:  21 (12 variable)
initial  value 124.143189
iter  10 value 36.581074
iter  20 value 34.686003
iter  30 value 34.676842
```

```
iter  40 value 34.671003
iter  50 value 34.670618
final  value 34.670617
converged

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

Coefficients:
  (Intercept) Calan.TreatmentDDG Calan.TreatmentMET     SEXSTR
2   -1.734529        -20.7937399        -1.15589609 -0.2803915
3  -19.374020          0.7406987        -0.05867808  5.9970139
  Calan.TreatmentDDG:SEXSTR Calan.TreatmentMET:SEXSTR
2                 -4.122079                 0.4628294
3                  9.545318                11.8261828

Std. Errors:
  (Intercept) Calan.TreatmentDDG Calan.TreatmentMET      SEXSTR
2   0.6262084       4.879631e-09           1.203222   0.9791894
3 103.8662689       1.036460e+02         103.645620 103.4249381
  Calan.TreatmentDDG:SEXSTR Calan.TreatmentMET:SEXSTR
2              1.697924e-09                  1.755282
3              1.036477e+02                103.646552

Residual Deviance: 69.34123
AIC: 93.34123
```

Pairwise Comparisons

```
SEX = HFR:
 contrast   estimate       SE df t.ratio p.value
 CON - DDG  5.14e-17      NaN 12     NaN     NaN
 CON - MET -1.17e-17      NaN 12     NaN     NaN
 DDG - MET -6.31e-17 2.44e-10 12   0.000  1.0000

SEX = STR:
 contrast   estimate       SE df t.ratio p.value
 CON - DDG  2.57e-18      NaN 12     NaN     NaN
 CON - MET -3.30e-17      NaN 12     NaN     NaN
 DDG - MET -3.56e-17 4.85e-10 12   0.000  1.0000

Results are averaged over the levels of: Calving.Ease
P value adjustment: tukey method for varying family sizes
```

Models for Calf Vigor

Ordinal Logistic Regression Model

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

Coefficients:
                            Value Std. Error t value
Calan.TreatmentDDG        -0.3742     0.8178 -0.4576
Calan.TreatmentMET        -0.1616     0.7599 -0.2127
SEXSTR                    -1.6731     1.1513 -1.4532
Calan.TreatmentDDG:SEXSTR  1.8571     1.4086  1.3184
Calan.TreatmentMET:SEXSTR  1.7870     1.3787  1.2961

Intercepts:
    Value   Std. Error t value
1|2  1.0761  0.5133     2.0964
2|3  1.7628  0.5425     3.2493
3|4  3.3313  0.7403     4.5000
4|5  3.7477  0.8448     4.4360

Residual Deviance: 160.1861
AIC: 178.1861

                                Value Std. Error    t value      p-value
Calan.TreatmentDDG        -0.3742104  0.8177783 -0.4575940 6.472442e-01
Calan.TreatmentMET        -0.1616211  0.7599288 -0.2126792 8.315772e-01
SEXSTR                    -1.6730691  1.1513128 -1.4531838 1.461727e-01
Calan.TreatmentDDG:SEXSTR  1.8570878  1.4086188  1.3183750 1.873782e-01
Calan.TreatmentMET:SEXSTR  1.7870376  1.3787308  1.2961468 1.949249e-01
1|2                        1.0760995  0.5133124  2.0963834 3.604819e-02
2|3                        1.7628214  0.5425236  3.2492990 1.156898e-03
3|4                        3.3312996  0.7402816  4.5000439 6.793943e-06
4|5                        3.7477448  0.8448396  4.4360433 9.162743e-06
```

Pairwise comparisons

```
SEX = HFR:
 contrast  estimate    SE  df z.ratio p.value
```

```
 CON - DDG    0.374 0.818 Inf   0.458  0.8910
 CON - MET    0.162 0.760 Inf   0.213  0.9754
 DDG - MET   -0.213 0.850 Inf  -0.250  0.9661

SEX = STR:
 contrast  estimate    SE  df z.ratio p.value
 CON - DDG   -1.483 1.147 Inf  -1.293  0.3990
 CON - MET   -1.625 1.150 Inf  -1.413  0.3342
 DDG - MET   -0.143 0.716 Inf  -0.199  0.9784

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

Multinomial Logistic Regression

```
# weights:  35 (24 variable)
initial  value 181.866484
iter  10 value 72.424368
iter  20 value 70.354245
iter  30 value 70.292462
iter  40 value 70.288076
final  value 70.287948
converged

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

Coefficients:
  (Intercept) Calan.TreatmentDDG Calan.TreatmentMET      SEXSTR
2   -2.014904          0.2231357        -38.7656488 -47.2209029
3   -2.014776         -0.4702127          0.6930633  -0.7578969
4  -37.507025         -7.3618177          4.6924230  13.9963447
5   -2.708032        -14.1380100        -42.7800640 -29.0708500
  Calan.TreatmentDDG:SEXSTR Calan.TreatmentMET:SEXSTR
2                 46.815435                86.7486656
3                  1.045754               -42.0187508
4                 27.982098                 0.2371585
5                  6.461777                71.9199389

Std. Errors:
  (Intercept) Calan.TreatmentDDG Calan.TreatmentMET     SEXSTR
2   0.7527781           1.072385          0.4331898   0.466635
3   0.7527356           1.284531          0.9398243   1.276397
```

```
4 483.2990336         483.298955        963.9381822 483.298693
5   1.0327936        1313.649758          0.4874010   0.487401
  Calan.TreatmentDDG:SEXSTR Calan.TreatmentMET:SEXSTR
2              6.674788e-01                 0.4331898
3              1.807792e+00                       NaN
4              4.832990e+02               963.9378422
5              2.989303e-07                 0.4874010

Residual Deviance: 140.5759
AIC: 188.5759
```

Models for Final Calf Weight

Linear Model

```
Call:
lm(formula = Final.Calf.BW ~ Calan.Treatment * SEX, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-311.20  -61.13    7.89   54.59  344.59

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1261.000     25.531  49.390   <2e-16 ***
Calan.TreatmentDDG         -17.800     39.000  -0.456   0.6490
Calan.TreatmentMET         -43.895     36.579  -1.200   0.2328
SEXSTR                      84.412     37.666   2.241   0.0271 *
Calan.TreatmentDDG:SEXSTR    3.519     53.429   0.066   0.9476
Calan.TreatmentMET:SEXSTR   30.115     52.830   0.570   0.5699
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.2 on 107 degrees of freedom
Multiple R-squared:  0.1645,    Adjusted R-squared:  0.1255
F-statistic: 4.214 on 5 and 107 DF,  p-value: 0.001553
```

Tukey comparison

```
     Simultaneous Tests for General Linear Hypotheses
```

```
Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Final.Calf.BW ~ Calan.Treatment * SEX, data = data2)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
DDG - CON == 0   -17.80      39.00  -0.456    0.892
MET - CON == 0   -43.89      36.58  -1.200    0.455
MET - DDG == 0   -26.09      39.44  -0.662    0.786
(Adjusted p values reported -- single-step method)
```

Linear Mixed Model

This will need to be redone in SAS if we stick with it.

```
Linear mixed model fit by REML ['lmerMod']
Formula: Final.Calf.BW ~ Calan.Treatment * SEX + (1 | Pen..) + (1 | Sire)
   Data: data2

REML criterion at convergence: 1329.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.91784 -0.55007 -0.02894  0.53915  2.74964

Random effects:
 Groups   Name        Variance Std.Dev.
 Sire     (Intercept)  1800     42.43
 Pen..    (Intercept)     0      0.00
 Residual             11711    108.22
Number of obs: 113, groups:  Sire, 6; Pen.., 4

Fixed effects:
                          Estimate Std. Error t value
(Intercept)               1266.813     30.104  42.081
Calan.TreatmentDDG         -23.290     37.427  -0.622
Calan.TreatmentMET         -36.938     34.934  -1.057
SEXSTR                      90.685     35.958   2.522
Calan.TreatmentDDG:SEXSTR    8.894     50.781   0.175
Calan.TreatmentMET:SEXSTR   20.064     50.729   0.396
```

```
Correlation of Fixed Effects:
            (Intr) Cl.TDDG Cl.TMET SEXSTR C.TDDG:
Cln.TrtmDDG -0.535
Cln.TrtmMET -0.564  0.462
SEXSTR      -0.548  0.445   0.479
C.TDDG:SEXS  0.393 -0.729  -0.335  -0.703
C.TMET:SEXS  0.397 -0.324  -0.695  -0.715  0.503
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

Tukey comparison

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lmer(formula = Final.Calf.BW ~ Calan.Treatment * SEX + (1 | Pen..) +
    (1 | Sire), data = data2)

Linear Hypotheses:
               Estimate Std. Error z value Pr(>|z|)
DDG - CON == 0   -23.29      37.43  -0.622    0.808
MET - CON == 0   -36.94      34.93  -1.057    0.540
MET - DDG == 0   -13.65      37.60  -0.363    0.930
(Adjusted p values reported -- single-step method)
```

Pariwise comparison

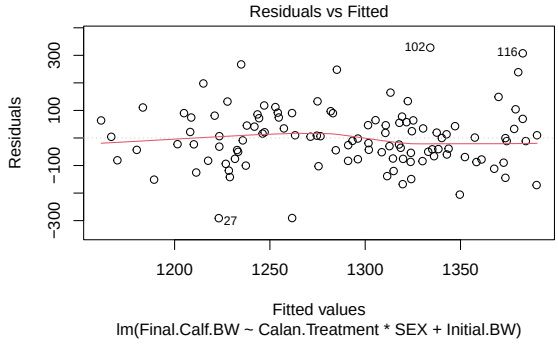
```
SEX = HFR:
 contrast  estimate   SE  df t.ratio p.value
 CON - DDG    23.29 38.1 104   0.612  0.8141
 CON - MET    36.94 35.4 103   1.043  0.5514
 DDG - MET    13.65 38.8 102   0.352  0.9340

SEX = STR:
 contrast  estimate   SE  df t.ratio p.value
 CON - DDG    14.40 35.1 102   0.410  0.9114
 CON - MET    16.87 36.8 102   0.459  0.8907
```

```
 DDG - MET     2.48 34.7 104   0.072  0.9972

Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 3 estimates
```

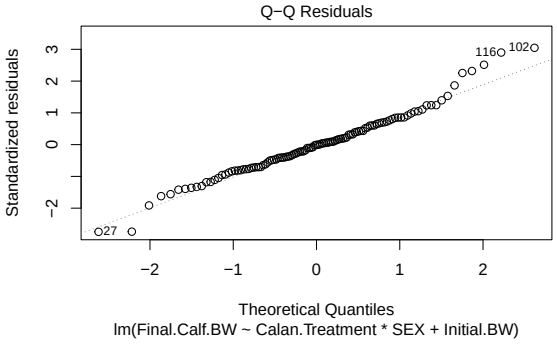
ANCOVA Model

This includes initial weight.

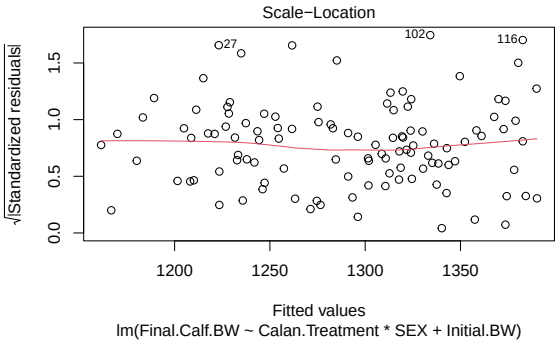
```
Call:
lm(formula = Final.Calf.BW ~ Calan.Treatment * SEX + Initial.BW,
    data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-291.16  -74.56   -0.57   63.79  327.87

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               717.7214   180.1893   3.983 0.000125 ***
Calan.TreatmentDDG        -22.9430    37.6139  -0.610 0.543193
Calan.TreatmentMET        -49.3793    35.2893  -1.399 0.164652
SEXSTR                     79.3296    36.3293   2.184 0.031194 *
Initial.BW                  0.5348     0.1757   3.044 0.002949 **
Calan.TreatmentDDG:SEXSTR  15.5936    51.6312   0.302 0.763230
Calan.TreatmentMET:SEXSTR  36.3226    50.9424   0.713 0.477404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110 on 106 degrees of freedom
Multiple R-squared:  0.2317,    Adjusted R-squared:  0.1882
F-statistic: 5.326 on 6 and 106 DF,  p-value: 7.747e-05
```



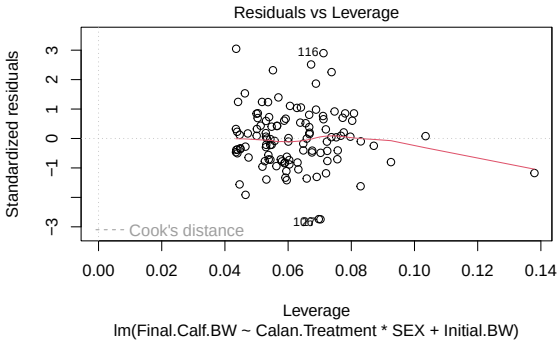
(a)



(b)



(c)



(d)

Post Hoc Test

```
SEX = HFR:
 contrast  estimate   SE  df t.ratio p.value
 CON - DDG    22.94 37.6 106   0.610  0.8150
 CON - MET    49.38 35.3 106   1.399  0.3449
 DDG - MET    26.44 38.0 106   0.696  0.7665

SEX = STR:
 contrast  estimate   SE  df t.ratio p.value
 CON - DDG     7.35 35.3 106   0.208  0.9763
 CON - MET    13.06 36.7 106   0.355  0.9328
 DDG - MET     5.71 34.2 106   0.167  0.9847

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

Model 3 - Binary Logistic Regression with GLM

```
Call:
glm(formula = Calf.Vigor.Binary ~ Calan.Treatment * SEX + Initial.BW,
    family = binomial(link = "logit"), data = data)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.5460511  5.1116605  -0.498    0.618
Calan.TreatmentDDG         -0.9116950  1.2106253  -0.753    0.451
Calan.TreatmentMET          0.4047963  0.8435442   0.480    0.631
SEXSTR                     -1.0458581  1.2072829  -0.866    0.386
Initial.BW                  0.0007978  0.0049818   0.160    0.873
Calan.TreatmentDDG:SEXSTR   1.7983853  1.7099003   1.052    0.293
Calan.TreatmentMET:SEXSTR  -0.5212389  1.6825267  -0.310    0.757

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80.667  on 112  degrees of freedom
Residual deviance: 77.041  on 106  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 91.041

Number of Fisher Scoring iterations: 5
```

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

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

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

Conclusion

Recomendation

References

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.

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

```
![](filename.png)
```