Data Analysis 3

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

November 18, 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.

RYAN SAYS NO IMPUTING CATEGORICAL VARS TOMA SAYS USE MODE - WE NEED TO PICK ONE

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.

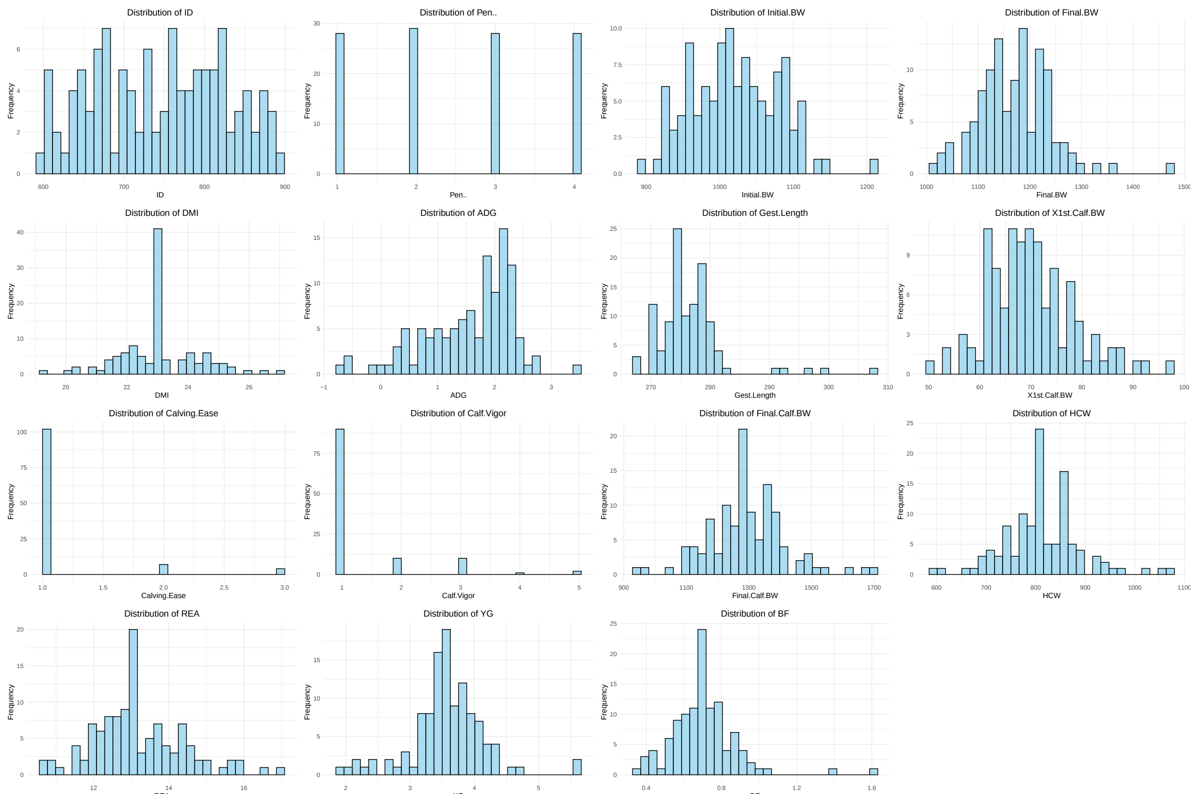
Summary Statistics

```
[1] "Summary Statistics for Numerical Variables"
```

```
   ID_mean    ID_sd ID_min ID_max ID_median Pen.._mean Pen.._sd Pen.._min
1 745.6549 80.98057    600    898       753   2.495575 1.119023         1
  Pen.._max Pen.._median Initial.BW_mean Initial.BW_sd Initial.BW_min
1         4            2        1020.947       59.4231            893
  Initial.BW_max Initial.BW_median Final.BW_mean Final.BW_sd Final.BW_min
1           1209              1017      1170.673    70.73207         1009
  Final.BW_max Final.BW_median DMI_mean   DMI_sd DMI_min DMI_max DMI_median
1         1468            1171 23.00597 1.291617   19.18   26.94     22.885
  ADG_mean    ADG_sd ADG_min ADG_max ADG_median Gest.Length_mean Gest.Length_sd
1 1.573097 0.7777487   -0.77     3.4       1.87         276.5487       5.639452
  Gest.Length_min Gest.Length_max Gest.Length_median X1st.Calf.BW_mean
1             267             307                276          70.09735
  X1st.Calf.BW_sd X1st.Calf.BW_min X1st.Calf.BW_max X1st.Calf.BW_median
1        8.459953               51               98                  69
  Calving.Ease_mean Calving.Ease_sd Calving.Ease_min Calving.Ease_max
1          1.132743       0.4331039                1                3
  Calving.Ease_median Calf.Vigor_mean Calf.Vigor_sd Calf.Vigor_min
1                   1        1.362832     0.8244256              1
  Calf.Vigor_max Calf.Vigor_median Final.Calf.BW_mean Final.Calf.BW_sd
1              5                 1           1290.106         122.0959
  Final.Calf.BW_min Final.Calf.BW_max Final.Calf.BW_median HCW_mean   HCW_sd
1               932              1690                 1283 812.7168 76.92859
  HCW_min HCW_max HCW_median REA_mean   REA_sd REA_min REA_max REA_median
1     587    1065        808  13.1694 1.205067   10.57   16.88      12.98
   YG_mean     YG_sd YG_min YG_max YG_median  BF_mean     BF_sd BF_min BF_max
1 3.576726 0.5783435   1.94   5.63      3.56 0.701885 0.1743249  0.334  1.596
  BF_median
1     0.681
```

```
[1] "Summary Statistics for Categorical Variables"
```

```
data frame with 0 columns and 1 row
```


Distribution of ID
ID
Distribution of Pen..
Pen..
Distribution of Initial.BW
Initial.BW
Distribution of Final.BW
Final.BW
Distribution of DMI
DMI
Distribution of ADG
ADG
Distribution of Gest.Length
Gest.Length
Distribution of X1st.Calf.BW
X1st.Calf.BW
Distribution of Calving.Ease
Calving.Ease
Distribution of Calf.Vigor
Calf.Vigor
Distribution of Final.Calf.BW
Final.Calf.BW
Distribution of HCW
HCW
Distribution of REA
REA
Distribution of YG
YG
Distribution of BF
BF
Frequency

	HFR	STR
CON	20	17
DDG	15	23
MET	19	19

Figure 2

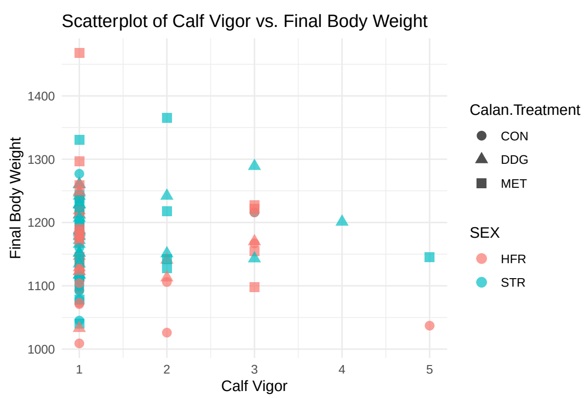


Figure 3

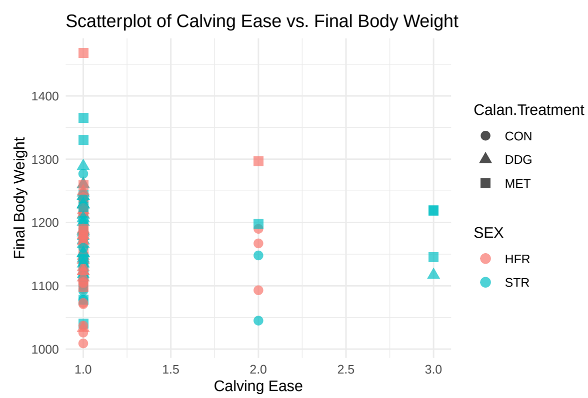


Figure 4

Exploring the Data

Potential models

Model 1 - ANCOVA Model

Post Hoc Test

```
SEX = HFR:
 contrast  estimate   SE  df t.ratio p.value
 CON - DDG    -31.4 21.5 106  -1.462  0.3131
 CON - MET    -59.6 20.1 106  -2.961  0.0105
 DDG - MET    -28.2 21.7 106  -1.302  0.3971

SEX = STR:
 contrast  estimate   SE  df t.ratio p.value
 CON - DDG    -50.1 20.1 106  -2.491  0.0377
 CON - MET    -39.9 20.9 106  -1.907  0.1417
 DDG - MET     10.2 19.5 106   0.521  0.8611

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

Model 2 - OLR Model

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

Coefficients:
                               Value Std. Error t value
Calan.TreatmentDDG        -0.3793001   0.714717 -0.5307
Calan.TreatmentMET        -0.1680591   0.669250 -0.2511
SEXSTR                    -1.6786727   0.423453 -3.9642
Initial.BW                 0.0004233   0.000447  0.9470
Calan.TreatmentDDG:SEXSTR  1.8668101   0.569521  3.2779
Calan.TreatmentMET:SEXSTR  1.7910506   0.583400  3.0700

Intercepts:
    Value   Std. Error t value
1|2  1.5044  0.0159    94.7233
2|3  2.1913  0.2098    10.4449
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-160.49  -32.56    1.86   41.27  195.29

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 690.3512   102.7726   6.717 9.52e-10 ***
Calan.TreatmentDDG           31.3751    21.4534   1.462  0.14657
Calan.TreatmentMET           59.5973    20.1276   2.961  0.00378 **
SEXSTR                       14.8792    20.7207   0.718  0.47429
Initial.BW                    0.4324     0.1002   4.314 3.60e-05 ***
Calan.TreatmentDDG:SEXSTR    18.7281    29.4483   0.636  0.52617
Calan.TreatmentMET:SEXSTR   -19.6531    29.0554  -0.676  0.50026
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.75 on 106 degrees of freedom
Multiple R-squared:  0.2552,    Adjusted R-squared:  0.2131
F-statistic: 6.054 on 6 and 106 DF,  p-value: 1.777e-05
```

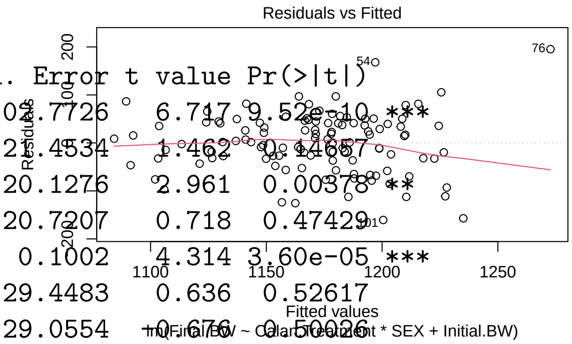


Figure 5

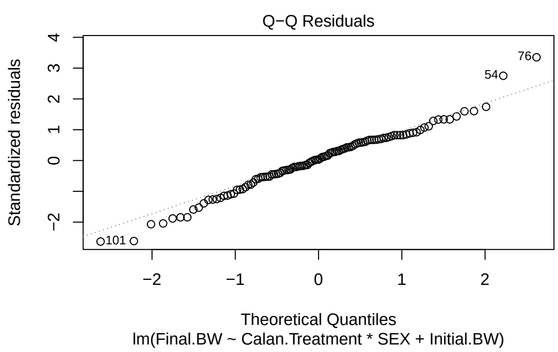


Figure 6

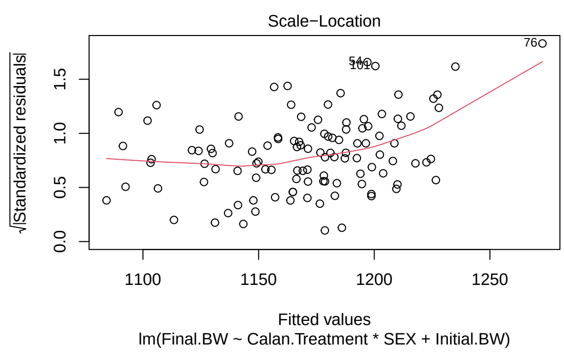


Figure 7

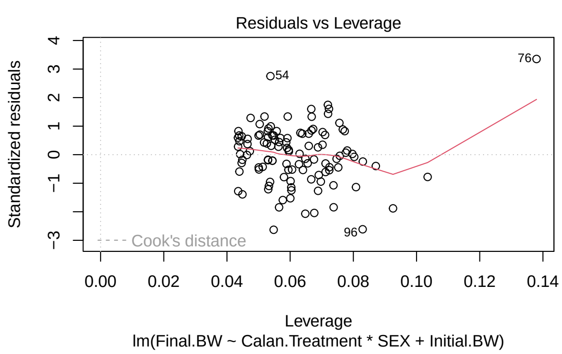


Figure 8

```
3|4  3.7598  0.5549     6.7752
4|5  4.1762  0.6890     6.0614

Residual Deviance: 160.1755
AIC: 180.1755
(7 observations deleted due to missingness)

                                  Value   Std. Error    t value      p-value
Calan.TreatmentDDG        -0.3793001124 0.7147168766 -0.5306998 5.956268e-01
Calan.TreatmentMET        -0.1680590719 0.6692497231 -0.2511156 8.017247e-01
SEXSTR                    -1.6786727385 0.4234529730 -3.9642483 7.362758e-05
Initial.BW                 0.0004232959 0.0004469635  0.9470481 3.436142e-01
Calan.TreatmentDDG:SEXSTR  1.8668100994 0.5695211115  3.2778593 1.045975e-03
Calan.TreatmentMET:SEXSTR  1.7910506485 0.5834004961  3.0700191 2.140451e-03
1|2                        1.5043654984 0.0158816886 94.7232713 0.000000e+00
2|3                        2.1912788200 0.2097933657 10.4449386 1.545554e-25
3|4                        3.7597956176 0.5549315376  6.7752423 1.241979e-11
4|5                        4.1761788945 0.6889770324  6.0614196 1.349253e-09
```

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.

Appendix A - R Code

Appendix B - SAS Code

Appendix C - Additional SAS Output

`![](filename.png)`