Initial EDA - 136 farms

A T Chamberlain

2024-07-24

# NOTES

# 24 July - increased from 105 to 136 herds - these are all routine HPRs from Emma

# 25 July - start sensible outlier checking - have removed 6+9 for BF + culling = 5%

## [1] "C:/ATC\_data/HPRs/Paper\_Eric\_OEE+LCA"

## Rows: 136 Columns: 51  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): Cull\_L6, Cull\_L6over  
## dbl (49): farmID, annualMY, herdSize, annualMYFPCM, AAFC, CI, culling, avLac...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## [1] "Herd number"

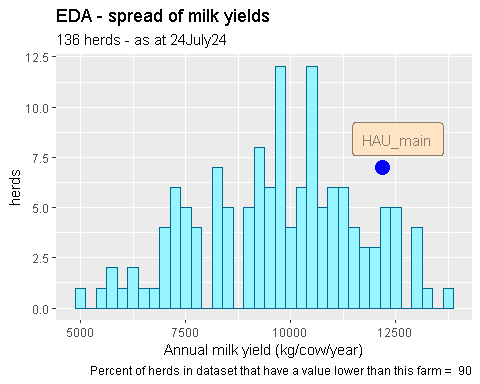
## [1] 1

## [1] "Herd Name"

## [1] "HAU\_main"

## [1] "Herd ID"

## [1] 56532401



## [1] "Value for this herd :"

## [1] 12182.2

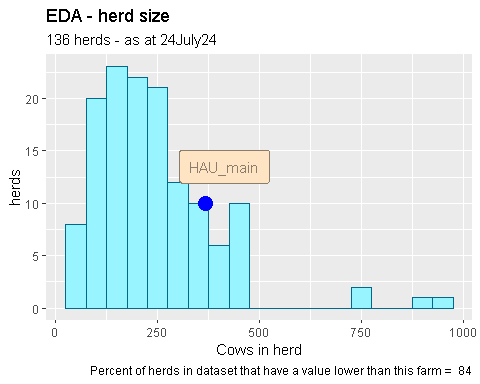
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5057 8355 9832 9722 10957 13653

## [1] 5056.66

## [1] 13653.26

## [1] 107974103

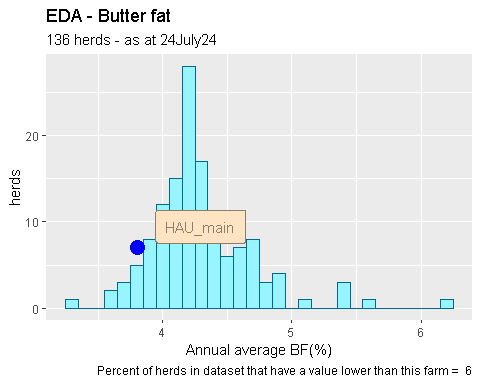
## [1] 42645



## [1] "Value for this herd :"

## [1] 366.67

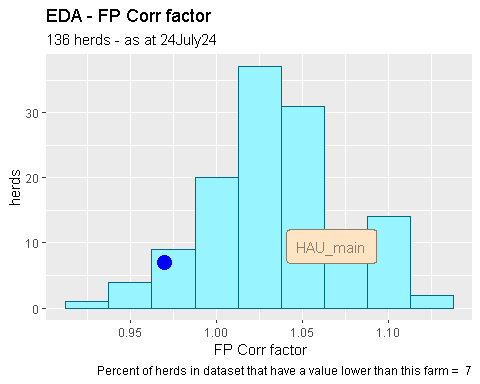
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 31.75 136.67 218.79 240.97 293.94 944.36



## [1] "Value for this herd :"

## [1] 3.801

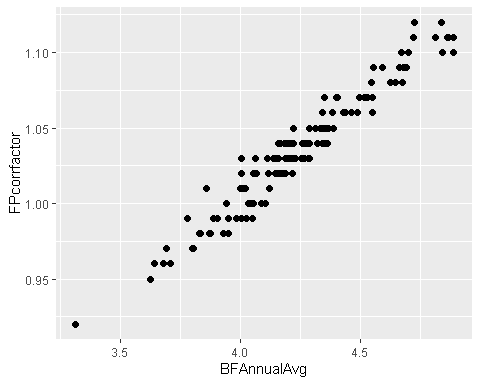
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.314 4.063 4.216 4.294 4.443 6.211



## [1] "Value for this herd :"

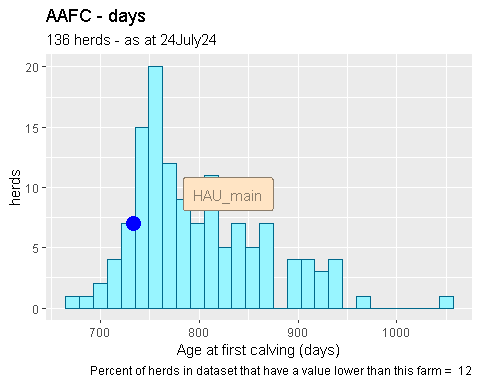
## [1] 0.97

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.920 1.010 1.030 1.034 1.060 1.120



## [1] 0.9695074

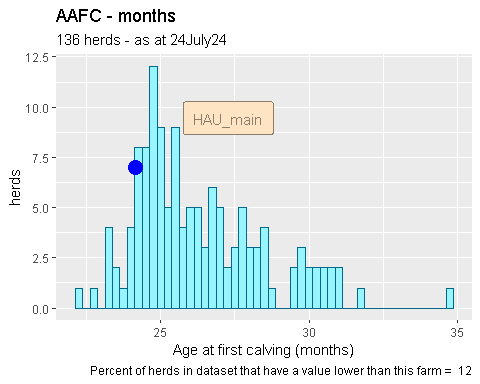
# look at distribution of the main driving KPI’s - excl MY - is any trimming needed?



## [1] "Value for this herd :"

## [1] 733.81

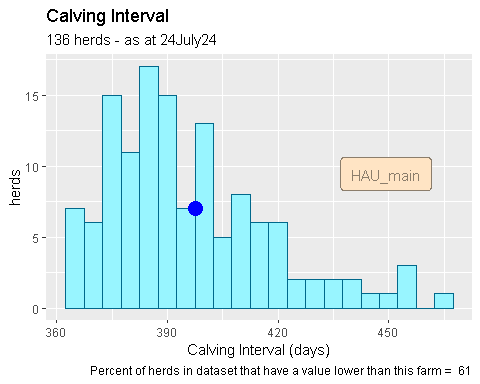
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 673.4 750.1 780.9 799.3 839.6 1055.2



## [1] "Value for this herd :"

## [1] 24.13849

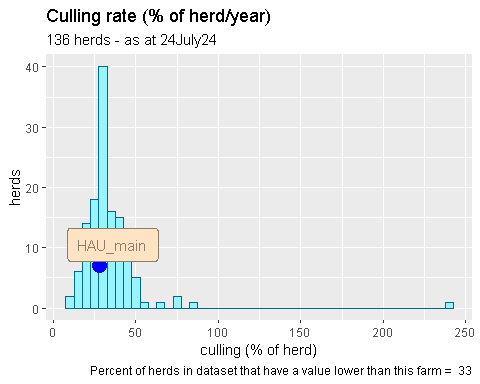
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 22.15 24.67 25.69 26.29 27.62 34.71



## [1] "Value for this herd :"

## [1] 397.58

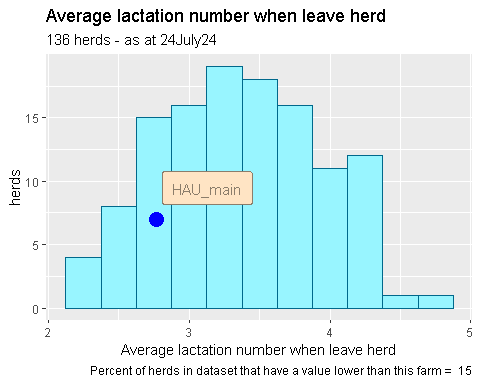
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 364.0 379.1 390.1 395.7 407.9 465.2



## [1] "Value for this herd :"

## [1] 27.74

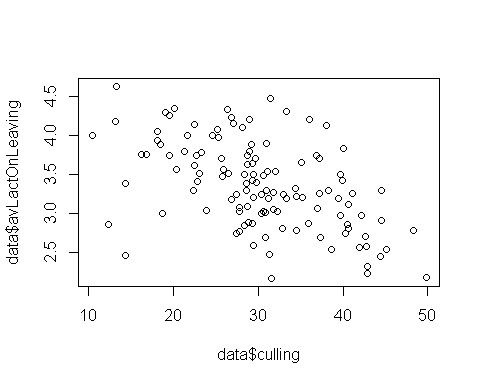
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.40 26.01 30.51 33.89 38.20 240.91



## [1] "Value for this herd :"

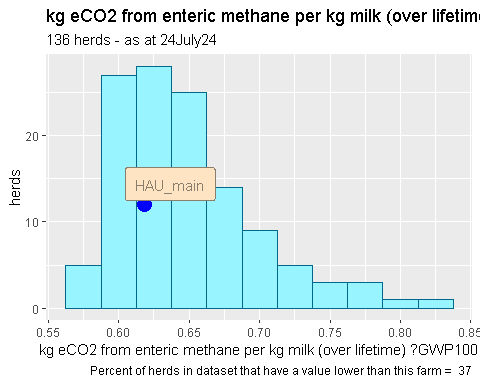
## [1] 2.77

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.170 2.980 3.300 3.365 3.760 4.630



## [1] -0.511474

# finally look at distribution of eCO2 / kg FP Corr milk



## [1] "Value for this herd :"

## [1] 0.618

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.5750 0.6120 0.6380 0.6482 0.6690 0.8230

SUMMARY - have deleted 6 herds as they looked like Channel Island breed and 9 herds as culling rate was too high due to TB culls, selling down-calved heifers, or downsizing / closing up / retiring.

Find out who are highest and lowest emitters

## [1] 0.575

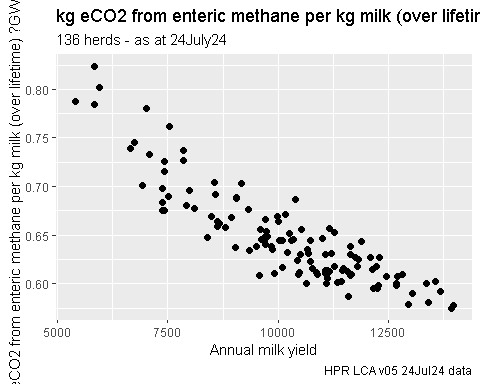
## [1] 0.823

## [1] 42645

## [1] 97468901

Finally - in EDA - look at some simple regressions

# milk yield   
  
p = ggplot(data, aes(x=annualMYFPCM, y=`eCO2 kg/kg FPCM`))+  
 geom\_point(size = 2) +  
 labs(x = "Annual milk yield",   
 y = "kg eCO2 from enteric methane per kg milk (over lifetime) ?GWP100",   
 title = "kg eCO2 from enteric methane per kg milk (over lifetime) cf milk yield",  
 subtitle = "136 herds - as at 24July24",  
 caption = "HPR LCA v05 24Jul24 data" )  
 # caption = result )   
  
plot(p)



corr = cor(data$annualMYFPCM,data$`eCO2 kg/kg FPCM`)  
  
print (corr)

## [1] -0.8850224

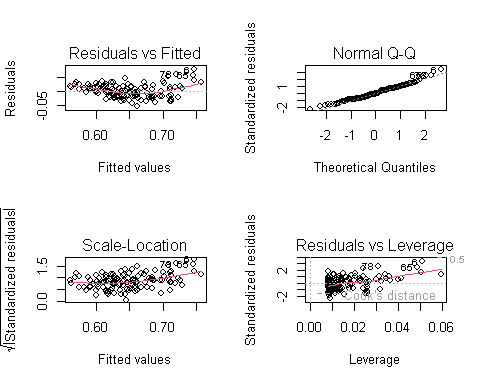
# regression   
  
model = lm(`eCO2 kg/kg FPCM`~ annualMYFPCM,data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = `eCO2 kg/kg FPCM` ~ annualMYFPCM, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.052516 -0.018039 -0.002677 0.015324 0.075094   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.816e-01 1.145e-02 76.99 <2e-16 \*\*\*  
## annualMYFPCM -2.295e-05 1.107e-06 -20.74 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02332 on 119 degrees of freedom  
## Multiple R-squared: 0.7833, Adjusted R-squared: 0.7814   
## F-statistic: 430.1 on 1 and 119 DF, p-value: < 2.2e-16

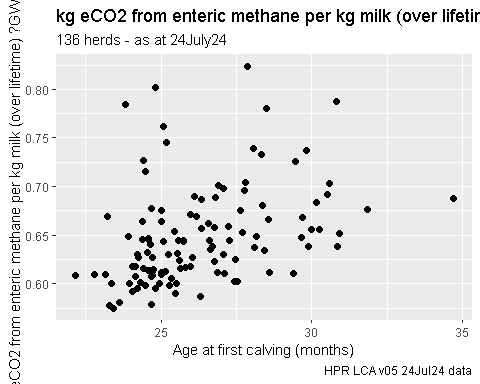
print (tidy(model))

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.882 0.0115 77.0 2.32e-103  
## 2 annualMYFPCM -0.0000230 0.00000111 -20.7 2.53e- 41

par(mfrow = c(2, 2))  
plot(model)



par(mfrow = c(1, 1))  
  
#Assumptions made when fitting the regression look valid  
  
# as it is such a strong effect (R2 = 0.75) should be be looking to model the residuals on the other factors   
  
# AAFC  
  
p = ggplot(data, aes(x=AAFCMonths, y=`eCO2 kg/kg FPCM`))+  
 geom\_point(size = 2) +  
 labs(x = "Age at first calving (months)",   
 y = "kg eCO2 from enteric methane per kg milk (over lifetime) ?GWP100",   
 title = "kg eCO2 from enteric methane per kg milk (over lifetime) cf milk yield",  
 subtitle = "136 herds - as at 24July24",  
 caption = "HPR LCA v05 24Jul24 data" )  
   
  
plot(p)



corr = cor(data$AAFCMonths,data$`eCO2 kg/kg FPCM`)  
  
print (corr)

## [1] 0.3845781

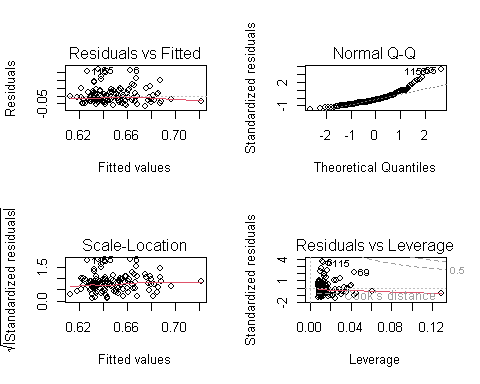
# regression - AAFC  
  
model = lm(`eCO2 kg/kg FPCM`~ AAFCMonths,data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = `eCO2 kg/kg FPCM` ~ AAFCMonths, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.06419 -0.02779 -0.01160 0.01456 0.16626   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.421936 0.049966 8.444 8.67e-14 \*\*\*  
## AAFCMonths 0.008621 0.001897 4.545 1.33e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04625 on 119 degrees of freedom  
## Multiple R-squared: 0.1479, Adjusted R-squared: 0.1407   
## F-statistic: 20.66 on 1 and 119 DF, p-value: 1.33e-05

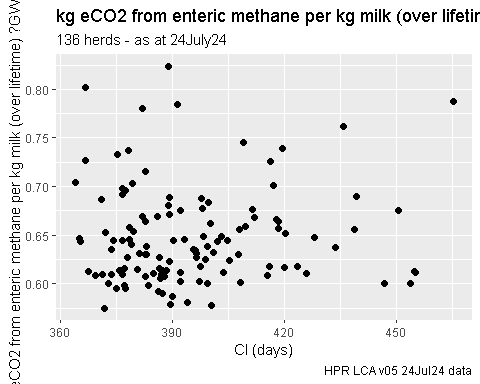
print (tidy(model))

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.422 0.0500 8.44 8.67e-14  
## 2 AAFCMonths 0.00862 0.00190 4.54 1.33e- 5

# AAFC has a significant effect but variation accounted for is small at 15%   
  
par(mfrow = c(2, 2))  
plot(model)



par(mfrow = c(1, 1))  
  
# CI days  
  
p = ggplot(data, aes(x=CI, y=`eCO2 kg/kg FPCM`))+  
 geom\_point(size = 2) +  
 labs(x = "CI (days)",   
 y = "kg eCO2 from enteric methane per kg milk (over lifetime) ?GWP100",   
 title = "kg eCO2 from enteric methane per kg milk (over lifetime) cf milk yield",  
 subtitle = "136 herds - as at 24July24",  
 caption = "HPR LCA v05 24Jul24 data" )  
   
  
plot(p)



corr = cor(data$CI,data$`eCO2 kg/kg FPCM`)  
  
print (corr)

## [1] 0.05935715

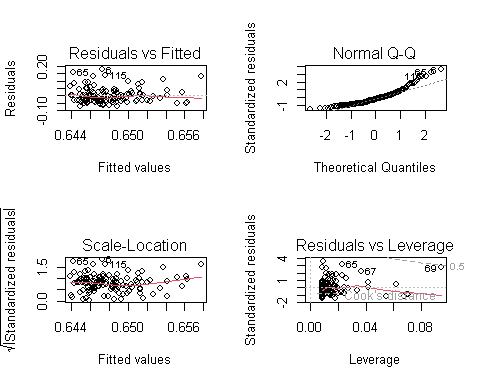
# regression   
  
model = lm(`eCO2 kg/kg FPCM`~ CI,data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = `eCO2 kg/kg FPCM` ~ CI, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.07089 -0.03565 -0.01074 0.02209 0.17570   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5943065 0.0832323 7.140 7.99e-11 \*\*\*  
## CI 0.0001363 0.0002101 0.649 0.518   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05001 on 119 degrees of freedom  
## Multiple R-squared: 0.003523, Adjusted R-squared: -0.00485   
## F-statistic: 0.4208 on 1 and 119 DF, p-value: 0.5178

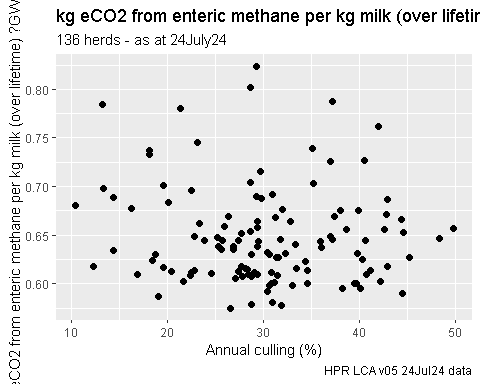
print (tidy(model))

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.594 0.0832 7.14 7.99e-11  
## 2 CI 0.000136 0.000210 0.649 5.18e- 1

par(mfrow = c(2, 2))  
plot(model)



par(mfrow = c(1, 1))  
  
# the slope - effect of CI - is not signif different from zero  
  
# culling   
  
p = ggplot(data, aes(x=culling, y=`eCO2 kg/kg FPCM`))+  
 geom\_point(size = 2) +  
 labs(x = "Annual culling (%)",   
 y = "kg eCO2 from enteric methane per kg milk (over lifetime) ?GWP100",   
 title = "kg eCO2 from enteric methane per kg milk (over lifetime) cf milk yield",  
 subtitle = "136 herds - as at 24July24",  
 caption = "HPR LCA v05 24Jul24 data" )  
   
  
plot(p)



corr = cor(data$culling,data$`eCO2 kg/kg FPCM`)  
  
print (corr)

## [1] -0.09784358

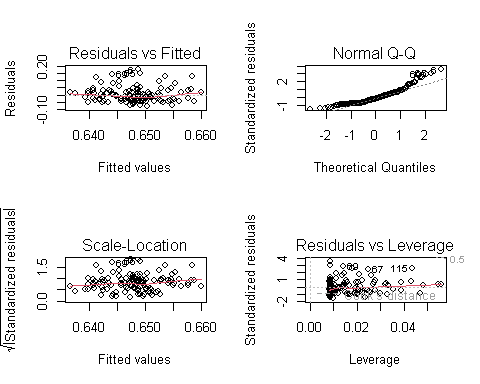
# regression   
  
model = lm(`eCO2 kg/kg FPCM`~ culling,data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = `eCO2 kg/kg FPCM` ~ culling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.07540 -0.03846 -0.01160 0.02112 0.17420   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6660006 0.0171918 38.739 <2e-16 \*\*\*  
## culling -0.0005886 0.0005489 -1.072 0.286   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04986 on 119 degrees of freedom  
## Multiple R-squared: 0.009573, Adjusted R-squared: 0.00125   
## F-statistic: 1.15 on 1 and 119 DF, p-value: 0.2857

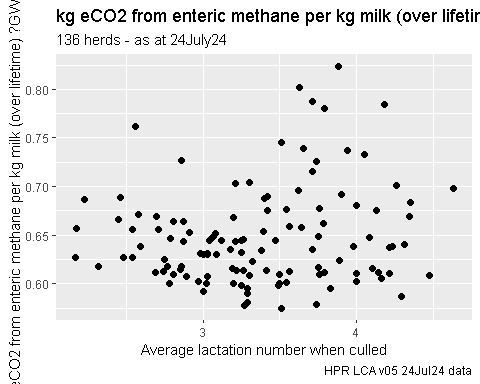
print (tidy(model))

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.666 0.0172 38.7 2.59e-69  
## 2 culling -0.000589 0.000549 -1.07 2.86e- 1

par(mfrow = c(2, 2))  
plot(model)



par(mfrow = c(1, 1))  
  
# not looking hopeful but need to remove the outlier(s) first   
  
# lact number at culling   
  
p = ggplot(data, aes(x=avLactOnLeaving, y=`eCO2 kg/kg FPCM`))+  
 geom\_point(size = 2) +  
 labs(x = "Average lactation number when culled",   
 y = "kg eCO2 from enteric methane per kg milk (over lifetime) ?GWP100",   
 title = "kg eCO2 from enteric methane per kg milk (over lifetime) cf milk yield",  
 subtitle = "136 herds - as at 24July24",  
 caption = "HPR LCA v05 24Jul24 data" )  
   
  
plot(p)



corr = cor(data$avLactOnLeaving,data$`eCO2 kg/kg FPCM`)  
  
print (corr)

## [1] 0.1454563

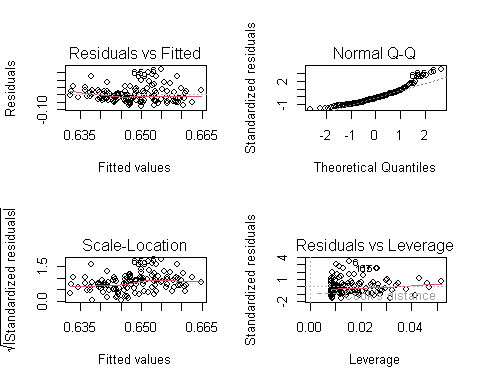
# multiple regression   
  
model = lm(`eCO2 kg/kg FPCM`~ + avLactOnLeaving,data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = `eCO2 kg/kg FPCM` ~ +avLactOnLeaving, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.07511 -0.03683 -0.01042 0.02362 0.16805   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.604182 0.027823 21.716 <2e-16 \*\*\*  
## avLactOnLeaving 0.013084 0.008158 1.604 0.111   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04957 on 119 degrees of freedom  
## Multiple R-squared: 0.02116, Adjusted R-squared: 0.01293   
## F-statistic: 2.572 on 1 and 119 DF, p-value: 0.1114

print (tidy(model))

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.604 0.0278 21.7 3.28e-43  
## 2 avLactOnLeaving 0.0131 0.00816 1.60 1.11e- 1

par(mfrow = c(2, 2))  
plot(model)



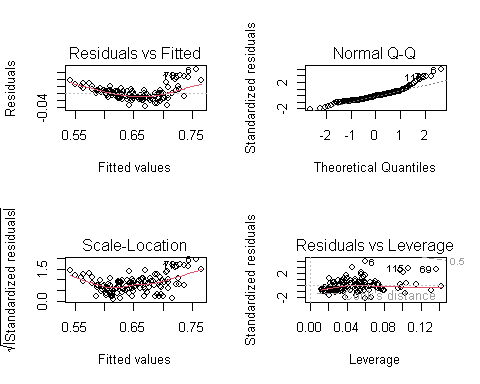
par(mfrow = c(1, 1))  
  
# finally all together   
  
model = lm(`eCO2 kg/kg FPCM`~ annualMYFPCM + AAFCMonths + CI + avLactOnLeaving + culling ,data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = `eCO2 kg/kg FPCM` ~ annualMYFPCM + AAFCMonths +   
## CI + avLactOnLeaving + culling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.035043 -0.012470 -0.001617 0.007178 0.066479   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.141e+00 4.332e-02 26.346 < 2e-16 \*\*\*  
## annualMYFPCM -2.670e-05 1.003e-06 -26.628 < 2e-16 \*\*\*  
## AAFCMonths 1.659e-03 7.920e-04 2.094 0.0384 \*   
## CI -4.689e-04 7.775e-05 -6.032 2.02e-08 \*\*\*  
## avLactOnLeaving -2.579e-02 3.595e-03 -7.173 7.65e-11 \*\*\*  
## culling 2.333e-04 2.267e-04 1.029 0.3056   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01749 on 115 degrees of freedom  
## Multiple R-squared: 0.8822, Adjusted R-squared: 0.8771   
## F-statistic: 172.2 on 5 and 115 DF, p-value: < 2.2e-16

print (tidy(model))

## # A tibble: 6 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 1.14 0.0433 26.3 1.52e-50  
## 2 annualMYFPCM -0.0000267 0.00000100 -26.6 5.31e-51  
## 3 AAFCMonths 0.00166 0.000792 2.09 3.84e- 2  
## 4 CI -0.000469 0.0000777 -6.03 2.02e- 8  
## 5 avLactOnLeaving -0.0258 0.00360 -7.17 7.65e-11  
## 6 culling 0.000233 0.000227 1.03 3.06e- 1

par(mfrow = c(2, 2))  
plot(model)



par(mfrow = c(1, 1))  
  
# now get a model that explains 87% of variance and all components are significant.   
  
# correlation matrix   
# select the variables wanted   
  
dataSelected = data[,c("farmID","eCO2 kg/kg FPCM","annualMY","annualMYFPCM","AAFCMonths","CI" , "avLactOnLeaving" , "culling")]  
  
  
  
res = cor(dataSelected)  
round(res, 2)

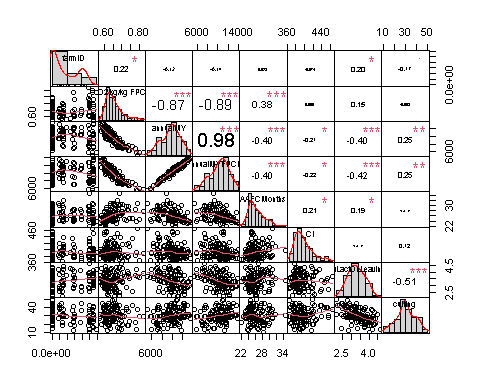
## farmID eCO2 kg/kg FPCM annualMY annualMYFPCM AAFCMonths CI  
## farmID 1.00 0.22 -0.12 -0.14 0.02 -0.07  
## eCO2 kg/kg FPCM 0.22 1.00 -0.87 -0.89 0.38 0.06  
## annualMY -0.12 -0.87 1.00 0.98 -0.40 -0.21  
## annualMYFPCM -0.14 -0.89 0.98 1.00 -0.40 -0.22  
## AAFCMonths 0.02 0.38 -0.40 -0.40 1.00 0.21  
## CI -0.07 0.06 -0.21 -0.22 0.21 1.00  
## avLactOnLeaving 0.20 0.15 -0.40 -0.42 0.19 -0.05  
## culling -0.17 -0.10 0.25 0.25 -0.05 0.12  
## avLactOnLeaving culling  
## farmID 0.20 -0.17  
## eCO2 kg/kg FPCM 0.15 -0.10  
## annualMY -0.40 0.25  
## annualMYFPCM -0.42 0.25  
## AAFCMonths 0.19 -0.05  
## CI -0.05 0.12  
## avLactOnLeaving 1.00 -0.51  
## culling -0.51 1.00

res2 = rcorr(as.matrix(dataSelected))  
res2

## farmID eCO2 kg/kg FPCM annualMY annualMYFPCM AAFCMonths CI  
## farmID 1.00 0.22 -0.12 -0.14 0.02 -0.07  
## eCO2 kg/kg FPCM 0.22 1.00 -0.87 -0.89 0.38 0.06  
## annualMY -0.12 -0.87 1.00 0.98 -0.40 -0.21  
## annualMYFPCM -0.14 -0.89 0.98 1.00 -0.40 -0.22  
## AAFCMonths 0.02 0.38 -0.40 -0.40 1.00 0.21  
## CI -0.07 0.06 -0.21 -0.22 0.21 1.00  
## avLactOnLeaving 0.20 0.15 -0.40 -0.42 0.19 -0.05  
## culling -0.17 -0.10 0.25 0.25 -0.05 0.12  
## avLactOnLeaving culling  
## farmID 0.20 -0.17  
## eCO2 kg/kg FPCM 0.15 -0.10  
## annualMY -0.40 0.25  
## annualMYFPCM -0.42 0.25  
## AAFCMonths 0.19 -0.05  
## CI -0.05 0.12  
## avLactOnLeaving 1.00 -0.51  
## culling -0.51 1.00  
##   
## n= 121   
##   
##   
## P  
## farmID eCO2 kg/kg FPCM annualMY annualMYFPCM AAFCMonths CI   
## farmID 0.0164 0.1883 0.1277 0.8064 0.4183  
## eCO2 kg/kg FPCM 0.0164 0.0000 0.0000 0.0000 0.5178  
## annualMY 0.1883 0.0000 0.0000 0.0000 0.0193  
## annualMYFPCM 0.1277 0.0000 0.0000 0.0000 0.0141  
## AAFCMonths 0.8064 0.0000 0.0000 0.0000 0.0207  
## CI 0.4183 0.5178 0.0193 0.0141 0.0207   
## avLactOnLeaving 0.0242 0.1114 0.0000 0.0000 0.0368 0.5955  
## culling 0.0679 0.2857 0.0065 0.0063 0.5955 0.1745  
## avLactOnLeaving culling  
## farmID 0.0242 0.0679   
## eCO2 kg/kg FPCM 0.1114 0.2857   
## annualMY 0.0000 0.0065   
## annualMYFPCM 0.0000 0.0063   
## AAFCMonths 0.0368 0.5955   
## CI 0.5955 0.1745   
## avLactOnLeaving 0.0000   
## culling 0.0000

p = chart.Correlation(dataSelected, histogram=TRUE, pch=19)

## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter



# finally save the data as a R dataset  
  
save(data,file = "C:/ATC\_data/HPRs/Paper\_Eric\_OEE+LCA/data\_25July24\_121herds")  
save(dataSelected,file = "C:/ATC\_data/HPRs/Paper\_Eric\_OEE+LCA/data\_25July24\_121herds\_selected")

This is not the same as sensitivity analysis which I would do with the model I have

. Find an ‘average herd’ or typical herd - how? .. normalise each variable and express as SND’s then look for the herd with the combined / average min absolute SND.

. set all variables into model at the average values . vary one at a time to the decile values

identify several ‘typical’ herds and do sens analysis on each herd and look at difference in results.