Longitudinal Analysis - Assignment I

124384 - Luycer Bosire

5/29/2021

We will use the following libraries in **R** to perform our analysis:

We import our data:

```
pksm_data <- read.csv("pksm data.csv")
dim(pksm_data)
```

```
## [1] 67 27
```

```
head(pksm_data,5)
```

```
##   county visit_no uniquefarmid uniquecalfid uniqobs hsize parity ageatvst
## 1      0        1            1            1       1    15      3       24
## 2      0        2            1            1       2    15      3       81
## 3      0        3            1            1       3    15      3      207
## 4      0        4            1            1       4    15      3      270
## 5      0        5            1            1       5    15      3      332
##   weight age1stdewm_wks strongyles coccidia breed freqmilkfed agewaterfree
## 1     65             24          0        0     0           3            4
## 2    143             24          0        0     0           3            4
## 3    182             24          0        0     0           3            4
## 4    246             24          0        0     0           3            4
## 5    276             24          0        0     0           3            4
##   agefedconcwks agefedhaywks weaned fecalconsist bcs socialtype floorraised
## 1             2            2      0            1 3.5          0           0
## 2             2            2      0            2 3.5          0           0
## 3             2            2      1            1 4.0          1           0
## 4             2            2      1            1 4.0          1           0
## 5             2            2      1            1 3.5          1           0
##   floorclean bedpresent tethered lying_day interaction
## 1          0          0        0        18           0
## 2          1          1        0        18           1
## 3          0          1        0        14           1
## 4          1          1        0        15           1
## 5          1          1        0        14           0
```

Next we perform data Wrangling

```
colnames(pksm_data)
```

```
##  [1] "county"         "visit_no"       "uniquefarmid"   "uniquecalfid"
##  [5] "uniqobs"        "hsize"          "parity"         "ageatvst"
##  [9] "weight"         "age1stdewm_wks" "strongyles"     "coccidia"
## [13] "breed"          "freqmilkfed"    "agewaterfree"   "agefedconcwks"
## [17] "agefedhaywks"   "weaned"         "fecalconsist"   "bcs"
```

```
## [21] "socialtype"     "floorraised"    "floorclean"     "bedpresent"
## [25] "tethered"       "lying_day"      "interaction"
```

```
pksm_data1 <- pksm_data %>%
  select(lying_day,strongyles,coccidia,breed,ageatvst,parity,weaned,floorraised,interaction)
dim(pksm_data1)
```

```
## [1] 67  9
```

```
str(pksm_data1)
```

```
## 'data.frame':    67 obs. of  9 variables:
##  $ lying_day  : int  18 18 14 15 14 13 18 13 13 14 ...
##  $ strongyles : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ coccidia   : int  0 0 0 0 0 200 0 100 200 400 ...
##  $ breed      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ageatvst   : int  24 81 207 270 332 388 32 158 221 283 ...
##  $ parity     : int  3 3 3 3 3 3 1 1 1 1 ...
##  $ weaned     : int  0 0 1 1 1 1 0 1 1 1 ...
##  $ floorraised: int  0 0 0 0 0 0 1 0 0 0 ...
##  $ interaction: int  0 1 1 1 0 0 1 1 1 1 ...
```

```
pksm_data1$weaned <- as.factor(pksm_data1$weaned)
pksm_data1$interaction <- as.factor(pksm_data1$interaction)
pksm_data1$breed <- as.factor(pksm_data1$breed)
pksm_data1$floorraised <- as.factor(pksm_data1$floorraised)

str(pksm_data1)
```

```
## 'data.frame':    67 obs. of  9 variables:
##  $ lying_day  : int  18 18 14 15 14 13 18 13 13 14 ...
##  $ strongyles : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ coccidia   : int  0 0 0 0 0 200 0 100 200 400 ...
##  $ breed      : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ageatvst   : int  24 81 207 270 332 388 32 158 221 283 ...
##  $ parity     : int  3 3 3 3 3 3 1 1 1 1 ...
##  $ weaned     : Factor w/ 2 levels "0","1": 1 1 2 2 2 2 1 2 2 2 ...
##  $ floorraised: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ interaction: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 2 2 2 2 ...
```

```
pksm_data1 %<>% mutate_if(is.integer,as.numeric)
str(pksm_data1)
```

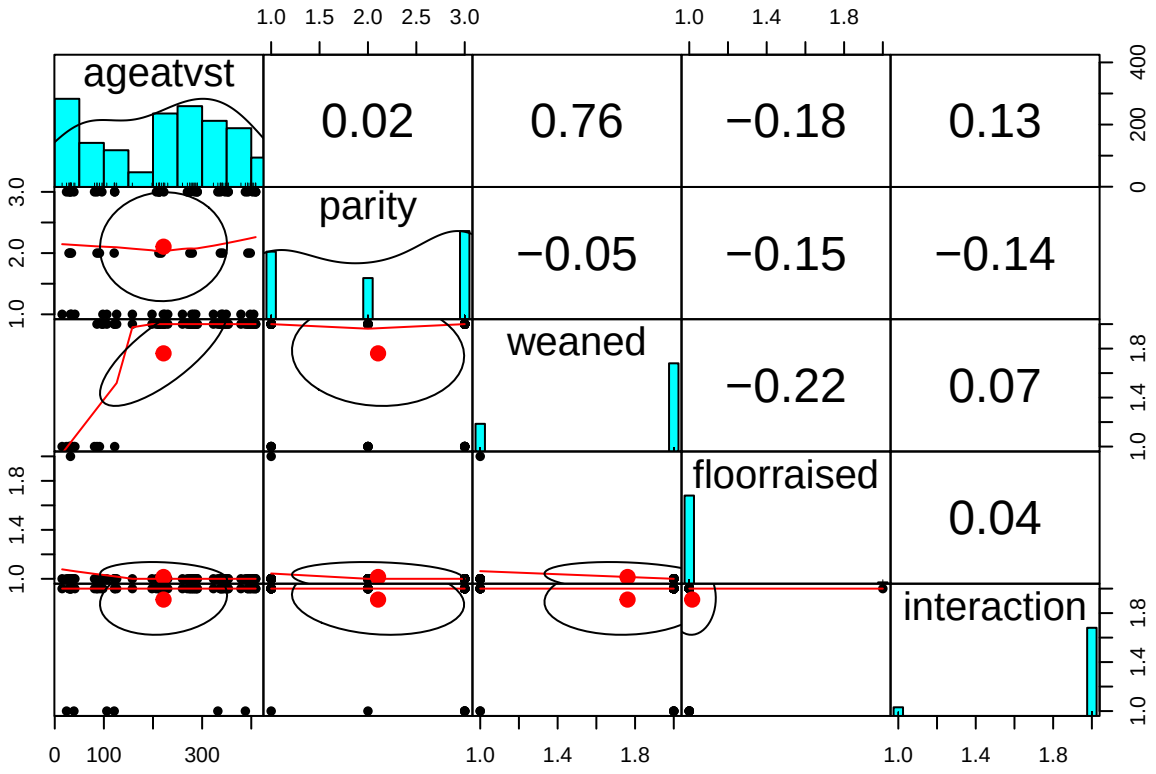
```
## 'data.frame':    67 obs. of  9 variables:
##  $ lying_day  : num  18 18 14 15 14 13 18 13 13 14 ...
##  $ strongyles : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ coccidia   : num  0 0 0 0 0 200 0 100 200 400 ...
##  $ breed      : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ageatvst   : num  24 81 207 270 332 388 32 158 221 283 ...
##  $ parity     : num  3 3 3 3 3 3 1 1 1 1 ...
##  $ weaned     : Factor w/ 2 levels "0","1": 1 1 2 2 2 2 1 2 2 2 ...
##  $ floorraised: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ interaction: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 2 2 2 2 ...
```

Removing the response variables and variables with zero fill

```
pksm_data2 <- subset(pksm_data1,select = -c(lying_day,strongyles,coccidia,breed))
```

We generate scatter plots to check for multi-correlation in the data

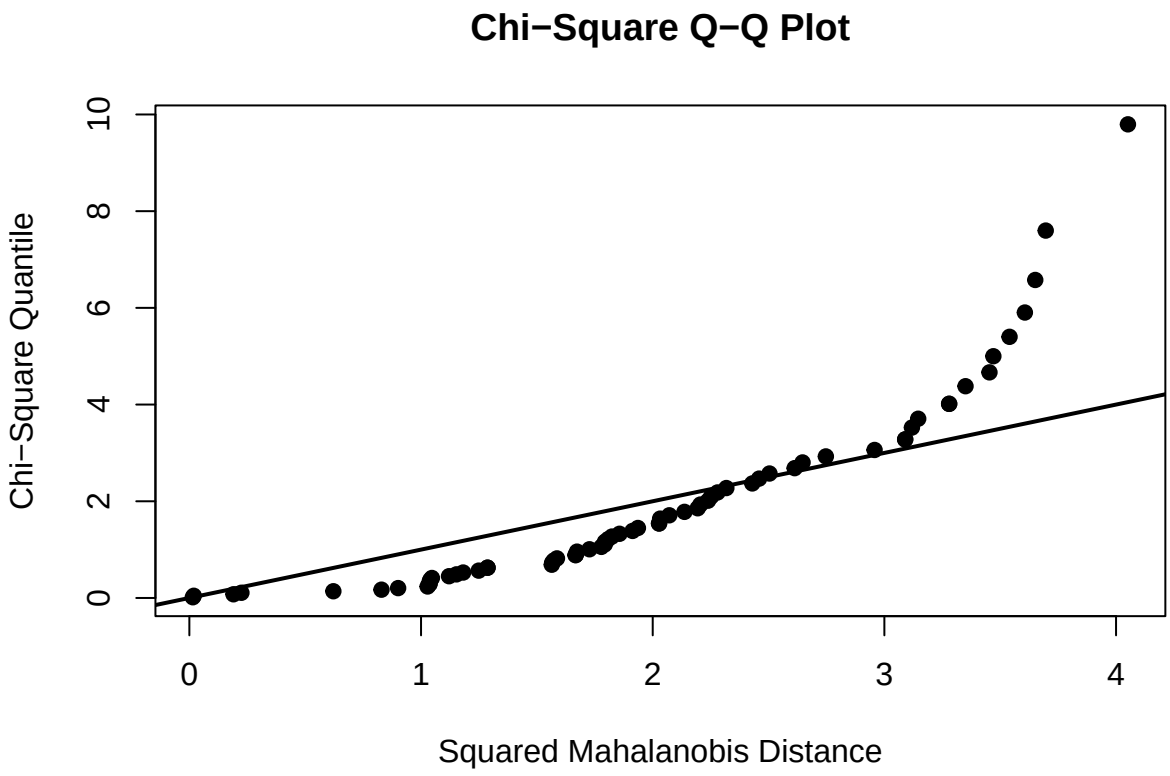
```
pairs.panels(pksm_data2,
             gap=0,
             pch = 16,
             cex=0.9)
```



There is a high correlation between *angeatvst* and *weaned* at 0.76. The other predictor variables have statistically insignificant correlation to induce multi-correlation as they have correlations less than 0.7.

We assess for Multivariate Collinearity:

```
pksm_data3 <- subset(pksm_data2,select = -c(weaned,floorraised,interaction))
mvn(pksm_data3,mvnTest = "mardia",scale = T,multivariatePlot = "qq")
```




```
## $multivariateNormality
##              Test         Statistic             p value Result
## 1 Mardia Skewness 0.987858098479116   0.911631475798832    YES
## 2 Mardia Kurtosis -3.14828155517961 0.00164233421818971     NO
## 3             MVN              <NA>                 <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk ageatvst     0.9161   2e-04      NO
## 2 Shapiro-Wilk  parity      0.7467  <0.001      NO
##
## $Descriptives
##           n       Mean     Std.Dev Median Min Max 25th 75th       Skew
## ageatvst 67 221.447761 129.3952465    224  15 409   97  338 -0.1801691
## parity   67   2.104478   0.8899154      2   1   3    1    3 -0.2003226
##           Kurtosis
## ageatvst -1.374055
## parity   -1.726317
```

The chi-square Q-Q plot indicates departures from multivariate normal distribution hence the data set is not Multivariate Normal.

We perform a Ridge Regression Using *lying_day* as the outcome variable:

```
pksm_data4 <- subset(pksm_data1,select = -c(strongyles,coccidia,breed))

set.seed(222)
ind <- sample(2,nrow(pksm_data4),replace=T,prob = c(0.7,0.3))
```

```
train <- pksm_data4[ind==1,]
test <- pksm_data4[ind==2,]

custom <- trainControl(method = "repeatedcv",
                       number = 10,
                      repeats=5,
                      verboseIter = F)
```

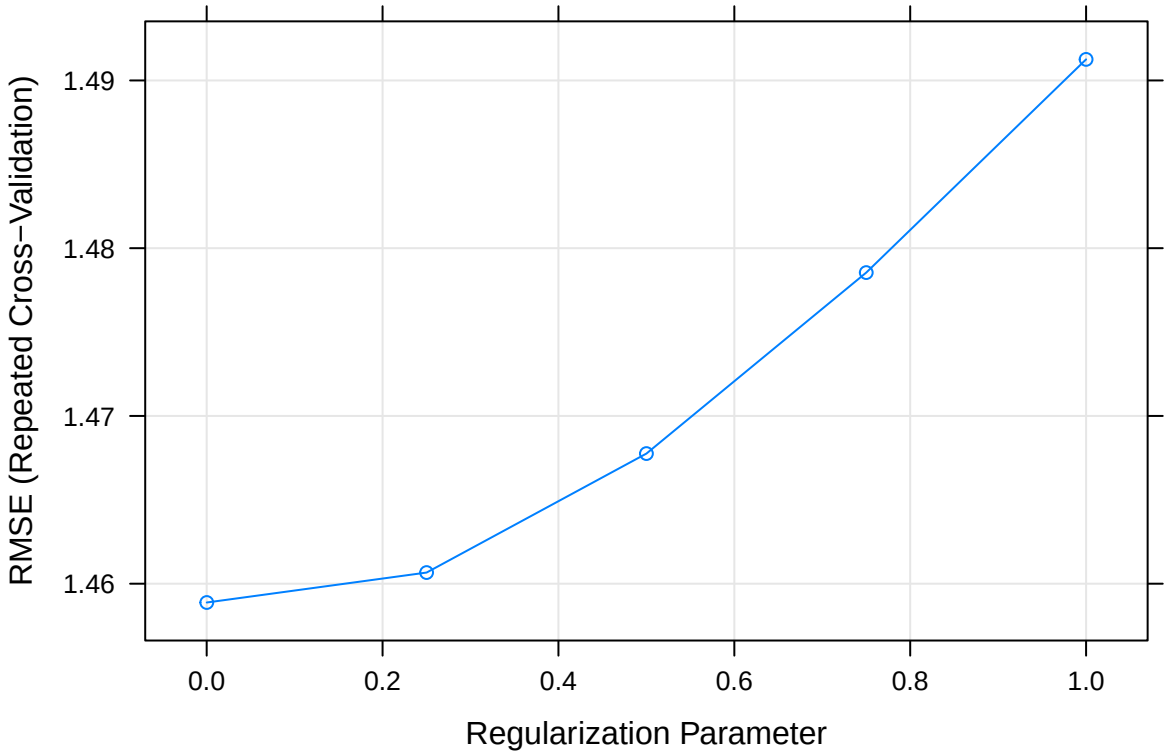
The Ridge Model

```
set.seed(1234)
ridge <- train(lying_day~.,data=pksm_data4,
               method='glmnet',
               tuneGrid=expand.grid(alpha=0,
               lambda=seq(0.0001,1,length=5)),
               trControl=custom)
ridge
```

```
## glmnet
##
## 67 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 60, 61, 60, 61, 61, 60, ...
## Resampling results across tuning parameters:
##
##   lambda    RMSE      Rsquared   MAE
##   0.000100  1.458879  0.5737612  1.205384
##   0.250075  1.460663  0.5742970  1.208074
##   0.500050  1.467757  0.5748649  1.213643
##   0.750025  1.478545  0.5751238  1.219888
##   1.000000  1.491256  0.5752932  1.227049
##
## Tuning parameter 'alpha' was held constant at a value of 0
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0 and lambda = 1e-04.
```

We plot the results and the coefficients of the model:

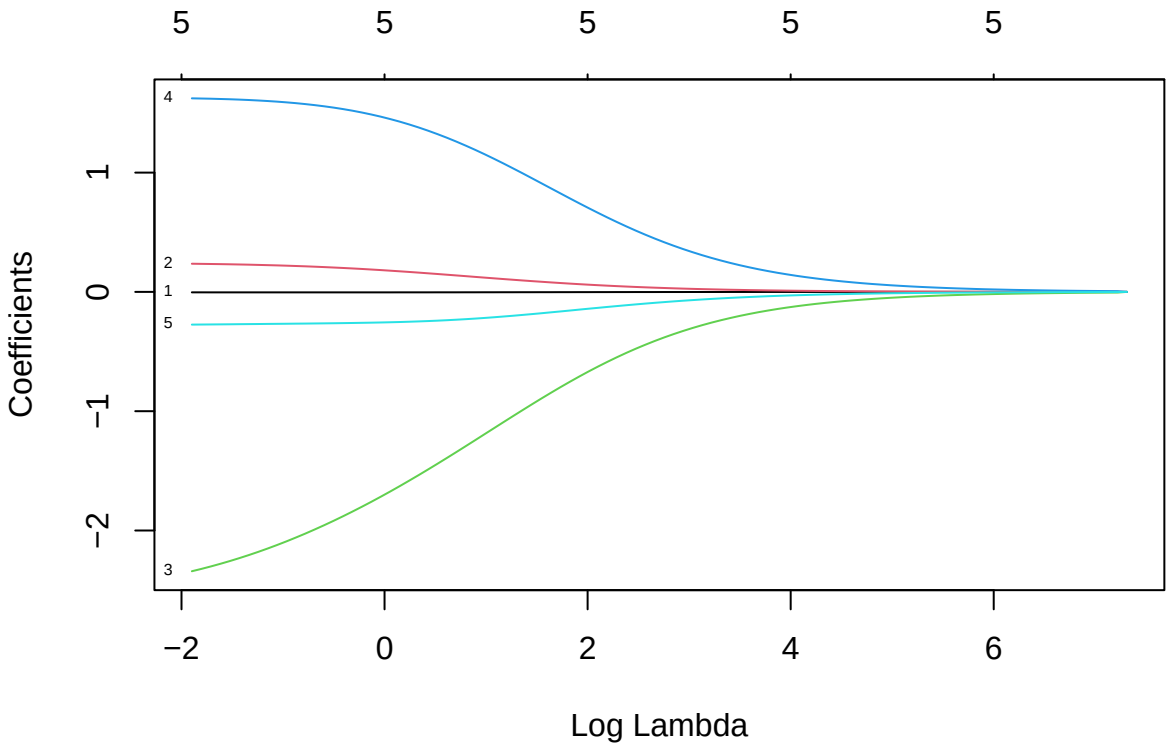
```
plot(ridge)
```




```
print(ridge)
```

```
## glmnet
##
## 67 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 60, 61, 60, 61, 61, 60, ...
## Resampling results across tuning parameters:
##
##   lambda    RMSE      Rsquared   MAE
##   0.000100  1.458879  0.5737612  1.205384
##   0.250075  1.460663  0.5742970  1.208074
##   0.500050  1.467757  0.5748649  1.213643
##   0.750025  1.478545  0.5751238  1.219888
##   1.000000  1.491256  0.5752932  1.227049
##
## Tuning parameter 'alpha' was held constant at a value of 0
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0 and lambda = 1e-04.
```

```
plot(ridge$finalModel,xvar = "lambda",label = T)
```

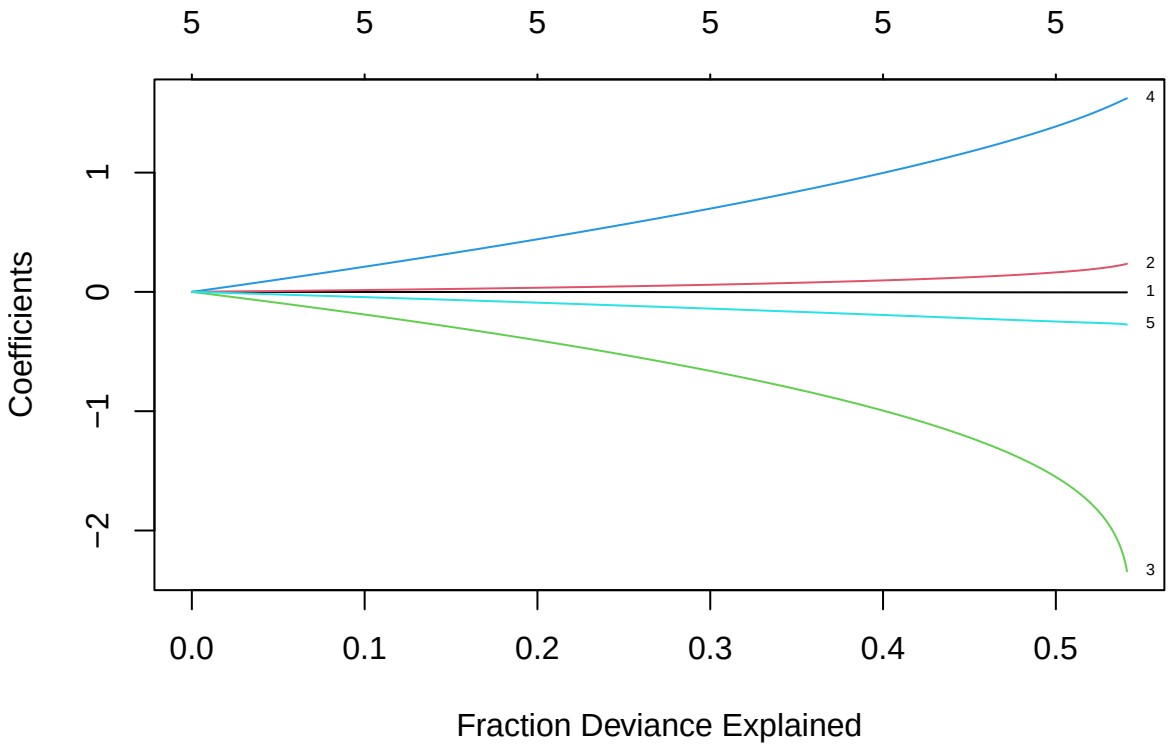


We can observe that for higher values of λ the RMSE increases. The RMSE sould be held lowest (1.458879) at the best value of lambda (1e-04).

For values of λ above 6, all the coefficients are more or less zero. As λ approaches zero, all the coefficients increase in value. At the top of the plot, it shows that we have all the 5 predictor variables in the model. This proves that indeed the Ridge does not shrink coefficients of the less significant variables to zero.

We perform another analysis by making the x variable the 'dev':

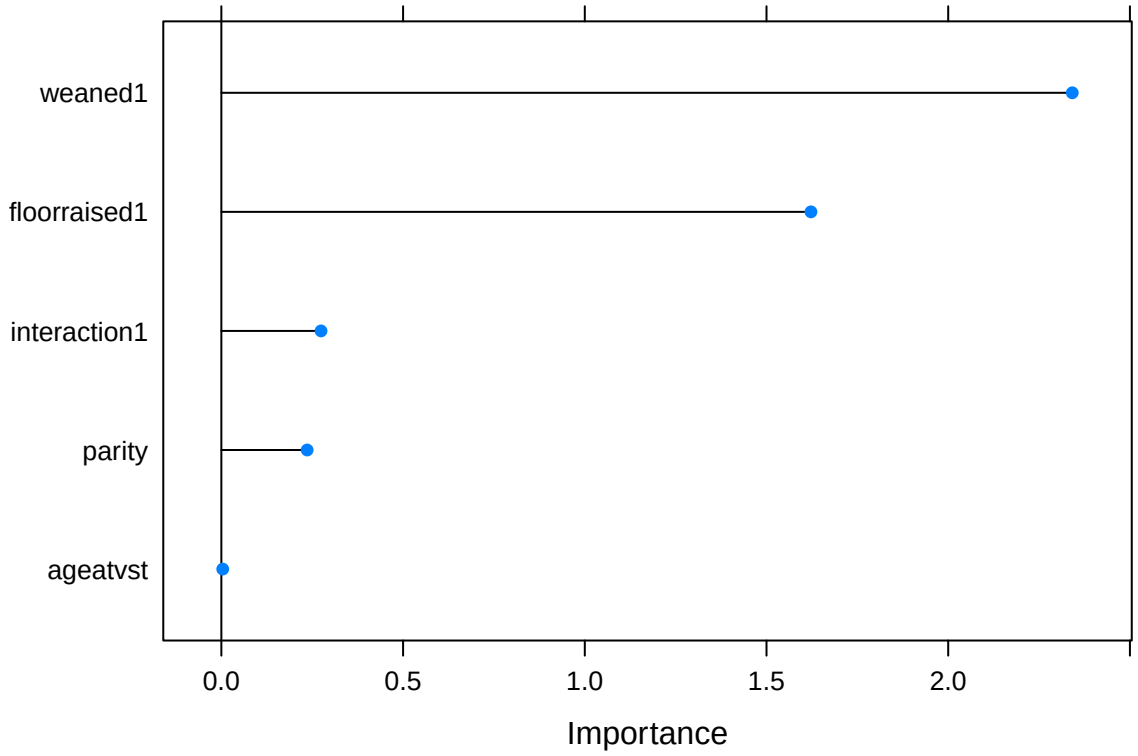
```
plot(ridge$finalModel,xvar = "dev",label = T)
```



This results on the x axis **Fraction Deviance Explained.** Beyond 40% of the deviance (variation), there is a sudden jump and the coefficients become highly inflated and therefore above 0.4 point on the deviance, over-fitting start to take place.

We assess variable importance plot: ***Variable Importance Plot - Scale = False***

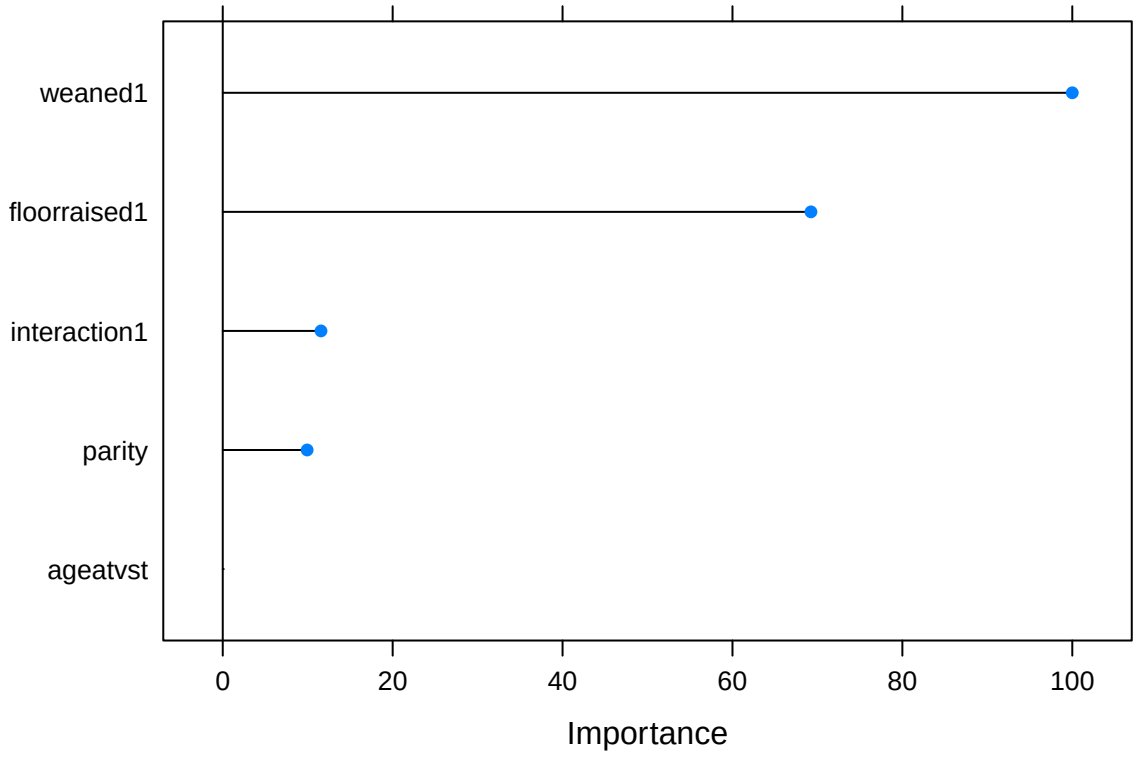
```
plot(varImp(ridge,scale = F))
```



weaned1 is the most important variable followed by *flooraised.* while the least ones are at the bottom, with low coefficient estimates.

Variable Importance Plot - Scale = True

```
plot(varImp(ridge,scale = T))
```



The scale has changed to be between 0 and 100

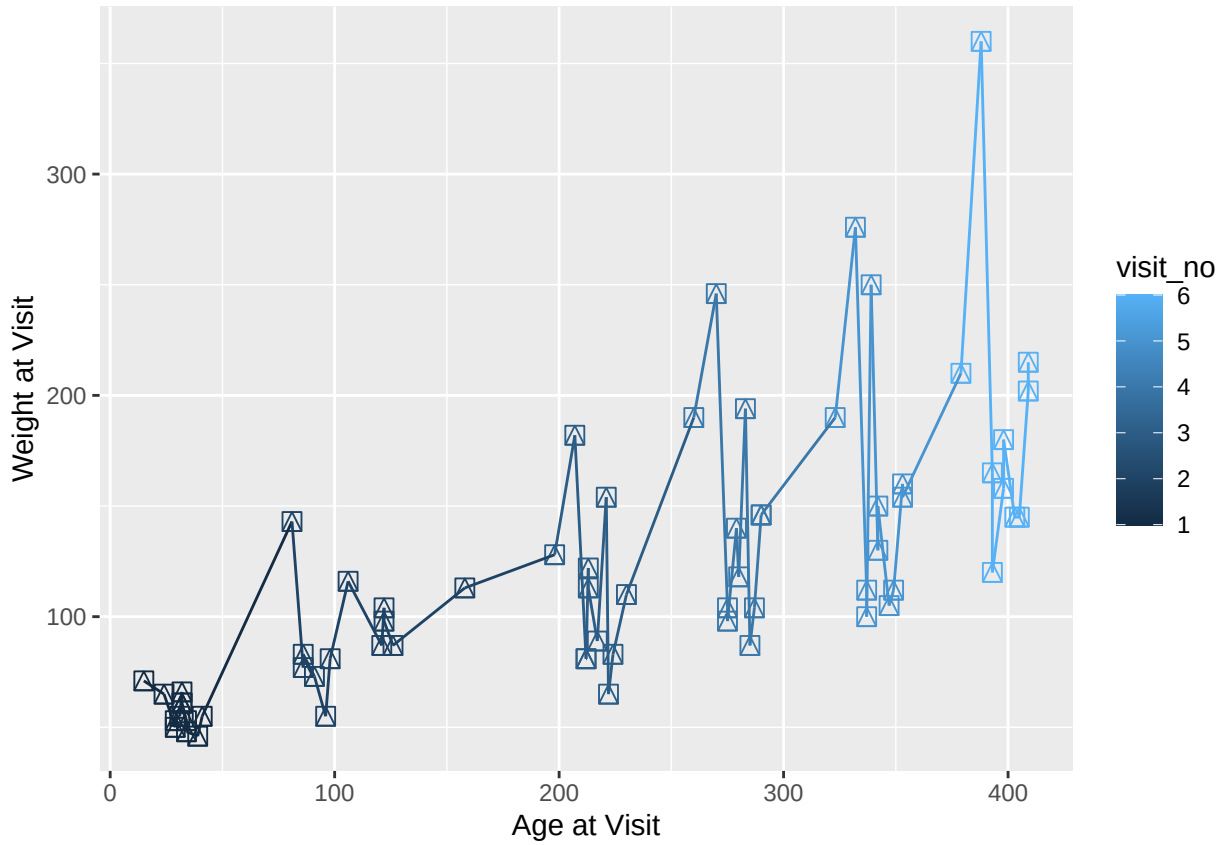
The OLS model

```
fit <- lm(lying_day~.,data = pksm_data4)
model_parameters(fit)
```

```
## Parameter       | Coefficient |       SE |         95% CI | t(61) |      p
## ------------------------------------------------------------------------------
## (Intercept)     |       16.52 |     0.87 | [14.78, 18.27] | 18.93 | < .001
## ageatvst        |   -3.42e-03 | 2.22e-03 | [-0.01,  0.00] | -1.54 | 0.128
## parity          |        0.24 |     0.21 | [-0.18,  0.67] |  1.15 | 0.254
## weaned [1]      |       -2.59 |     0.67 | [-3.93, -1.25] | -3.86 | < .001
## floorraised [1] |        1.63 |     1.56 | [-1.50,  4.75] |  1.04 | 0.302
## interaction [1] |       -0.29 |     0.65 | [-1.59,  1.02] | -0.44 | 0.663
```

The Graph of Weight against Age at Visit given the Number of Visits:

```
ggplot(data=pksm_data,aes(x=ageatvst,y=weight,colour=visit_no))+geom_point(size=3,shape=14)+geom_line()+
```



From the graph, it is evident that both weight and age increase by the number of visits.