Tema Repaso: Modelos Lineales con R

@umh1465: Análisis estadístico de series económicas

Xavi Barber

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

2018-04-20

Revisando Conceptos

La relación entre variables

- Una de las principales herramientas estadísticas es la de intentar "entender" y explicar el comportamiento de las cosas mediante la recolección de la menor información posible.
- Esa información está formada por Variables Aleatorias, donde unas explicarán las otras.

Descriptivo

- Un descriptivo inicial nos dará información al respecto de dichas relaciones.
- A veces, también podremos necesitar de pruebas o cálculos que nos cuantifiquen esas posibles relaciones entre las Variables a estudio.

Modelos

- Queremos Explicar la relación existente entre Una variable, la cual creemos dependiente, y otras continuas o categóricas:

Regresión Lineal Simple

$$Y \sim X_1$$

Regresión Lineal Multiple

$$Y \sim X_1 + X_2 + \cdots + X_p$$

ANOVA

$$Y \sim F_1 + F_2 + \cdots + F_p$$

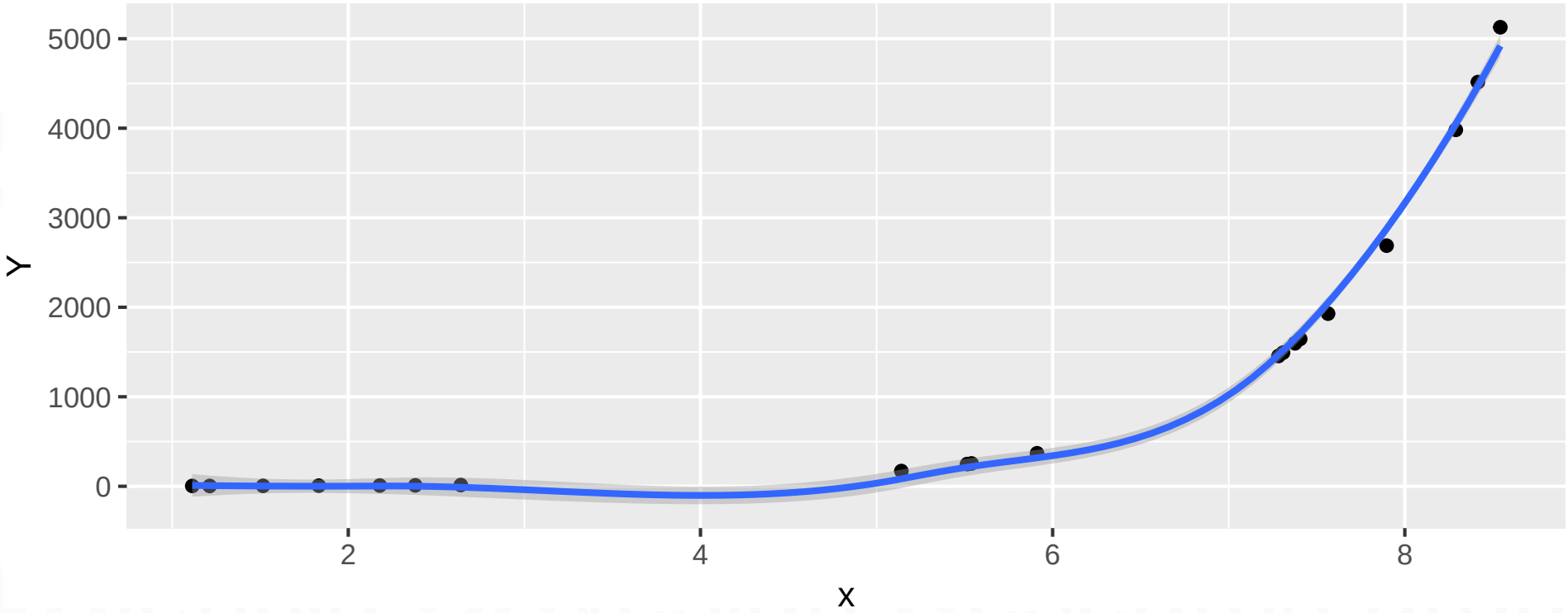
ANCOVA

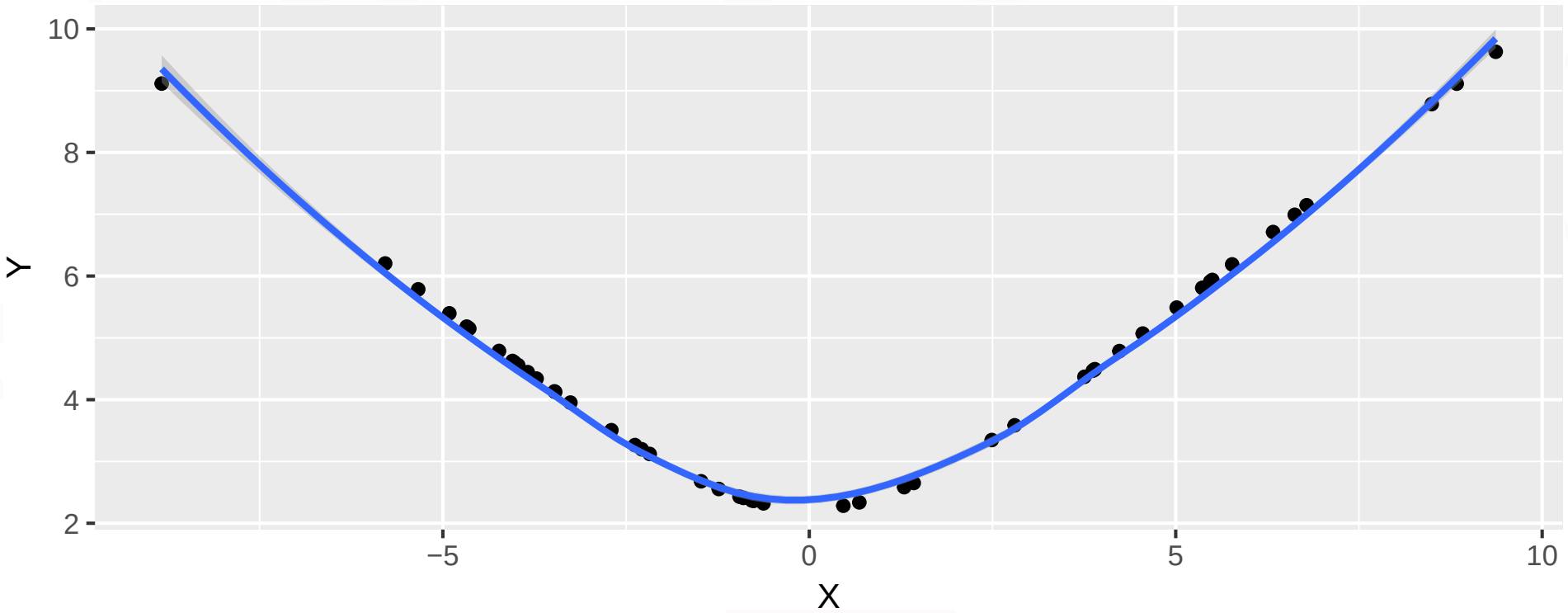
$$Y \sim X_1 + X_2 + \cdots + X_p + F_1 + F_2 + \cdots + F_p$$

Tanto en el ANOVA como en el ANCOVA pueden aparecer interacciones, es decir, que exista un comportamiento distinto al cruzar ciertos factores o facotores con variables.

Linealidad

- No es lo mismo decir que dos variables están relacionadas, que exista una Correlación Lineal entre las variables.




10
8
6
4
2
Y
−5
0
5
10
X

Linealidad vs Correlación

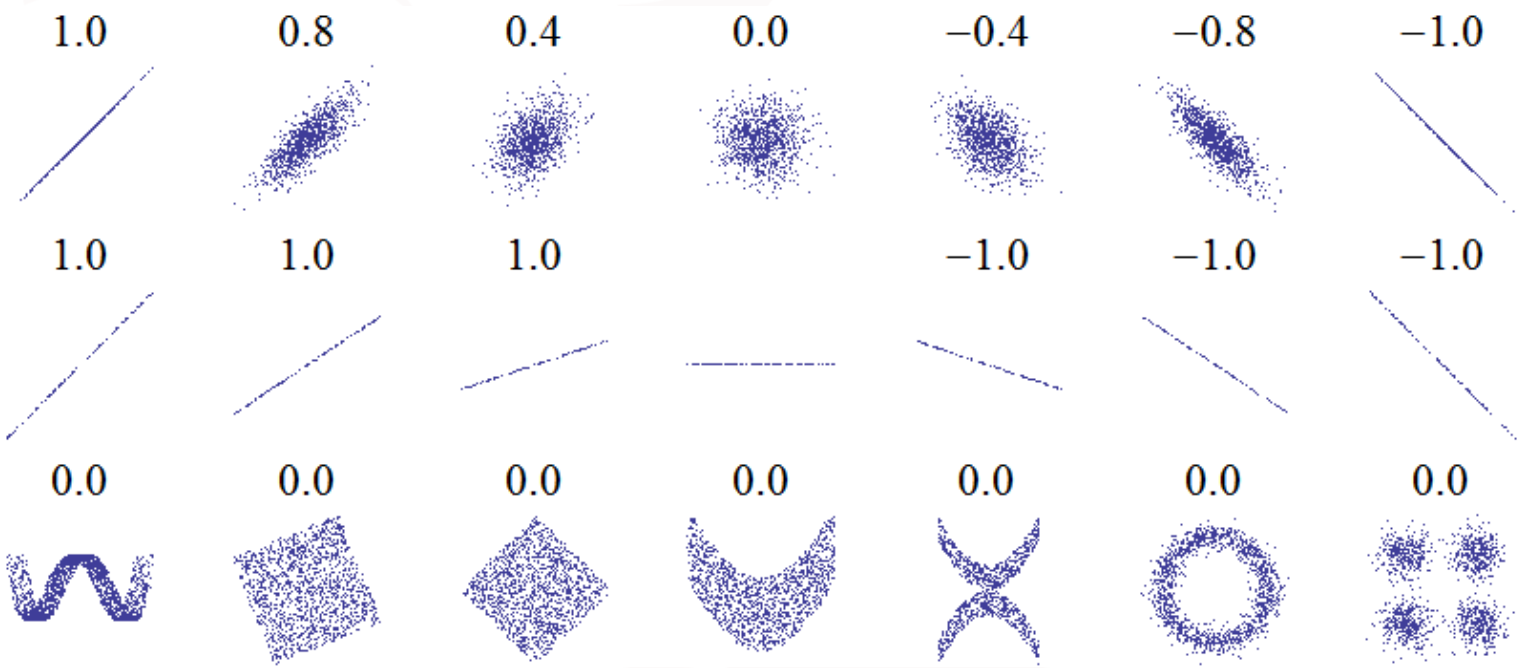


Figure 1

Linealidad

¿Existe relación Lineal entre las variables?

- Gráficamente y Numéricamente
- Si no aplicar las transformaciones habituales:
  - $log(Y) \sim X$;
  - $Y \sim log(X)$;
  - $log(Y) \sim log(X)$;
  - $\frac{1}{Y} \sim \frac{1}{X}$

Independencia

- ¿Son INDEPENDIENTES las X_i?
  - Gráficamente y Numéricamente
  - Eliminar las relaciones redundantes
  - Si existen excesivas X:
    - Aplicar un AF o CP para reducir la dimensión del problema.

Ejemplo

```
library(BCA)
data(Eggs)
```

The Eggs data set has 105 observations and 10 variables. The data contains information on weekly sales of eggs in Southern California over a two year period.

Week: The observation week (1 to 105). This variable can be used as a time trend. **Month**:A factor that gives the name of the month in which the observation occured.

First.Week:A factor indicating whether the observation fell on the first week of the month with levels: No Yes

Easter: A factor that indicates whether the observation fell the week prior to the week containing Easter Sunday, the week containing Easter Sunday, the week following the week containing Easter Sunday, or a non-Easter week with levels: Non Easter; Pre Easter; Easter; Post Easter.

Cases: Retail sales of eggs in cases.

Egg.Pr: Average retail egg price in cents per dozen.

Beef.Pr: Average retail price of 7-bone beef roast in cents per pound.

Pork.Pr: Average retail price of strip bacon in cents per pound.

Chicken.Pr:Average retail price of whole frying chicken in cents per pound.

Cereal.Pr: Average retail price of Cheerios breakfast cereal in cents per pound.

Source: Putler (1992)

Estudio Analítico y Gráfico de las Variables

- Debemso hacer un estudio pormenorizado de las variables , en concreto debemos tener en cuenta lo siguiente:
  - Descriptivo numérico centrándonos en Mínimos y Máximos, NAs, diferencias entre Media y Mediana, así como buscando desviaciones muy grandes respecto a la media.
  - Gráfico de Dispersión entre las contínuas.
  - Gráfico de Cajas respecto a lso factores con la variable respuesta.

Correlación entre var. independientes

El paquete **corrplot** ofece herramientas muy buenas para poder detectar de una forma gráfica las posibles variables que estén "linealmente" relacionadas.

- Primero deberemos seleccionar las variables de las cuales queremos obtener la correlación.
- **NUNCA** calculemos la correlación de variables tipo *factor*.

```
library(corrplot)
sel <- c(1, 5, 6, 7, 8, 9, 10)
datos2 <- data.frame(Eggs[, sel])
M <- cor(datos2)
corrplot(M, method = "number")
```

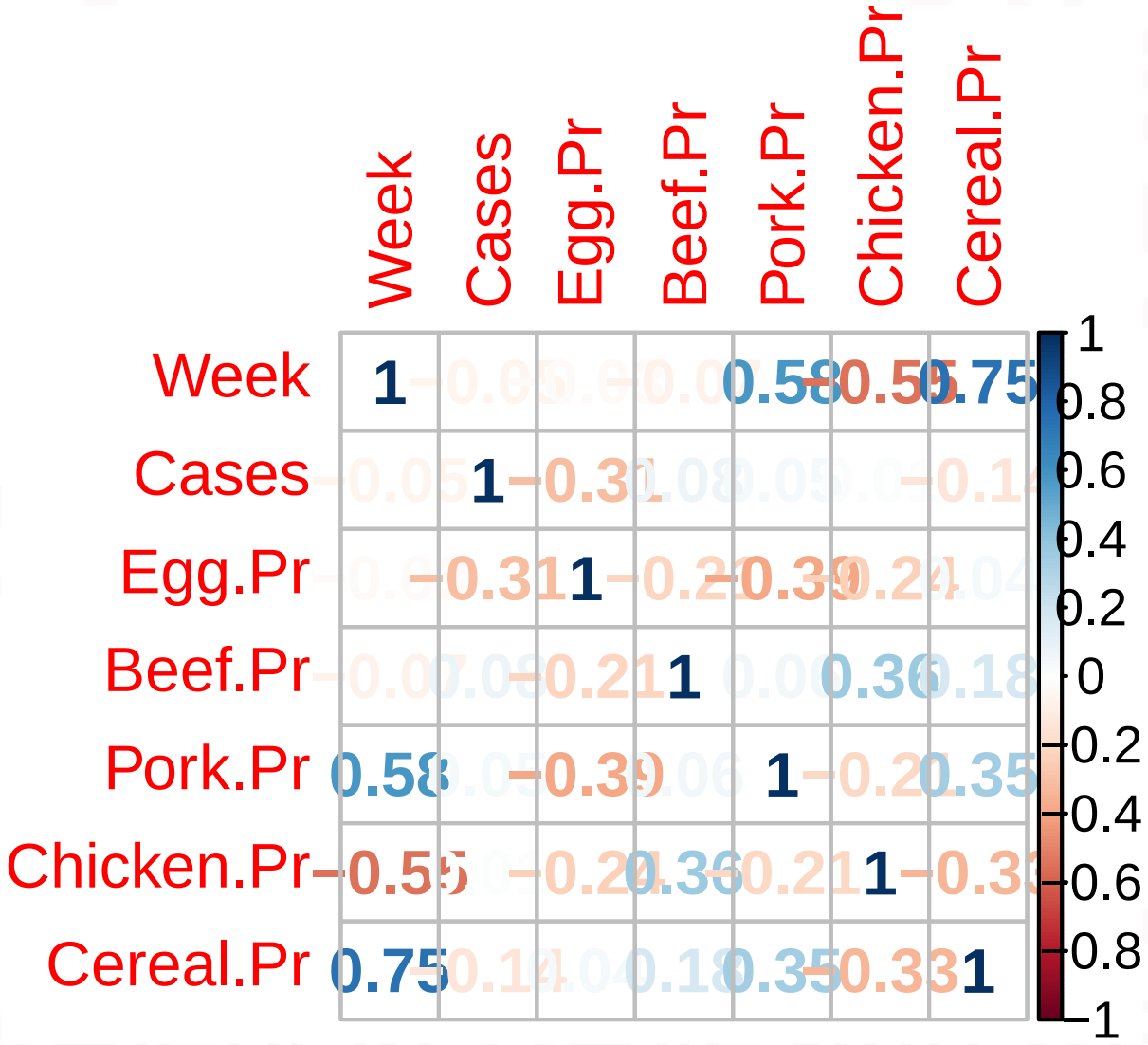
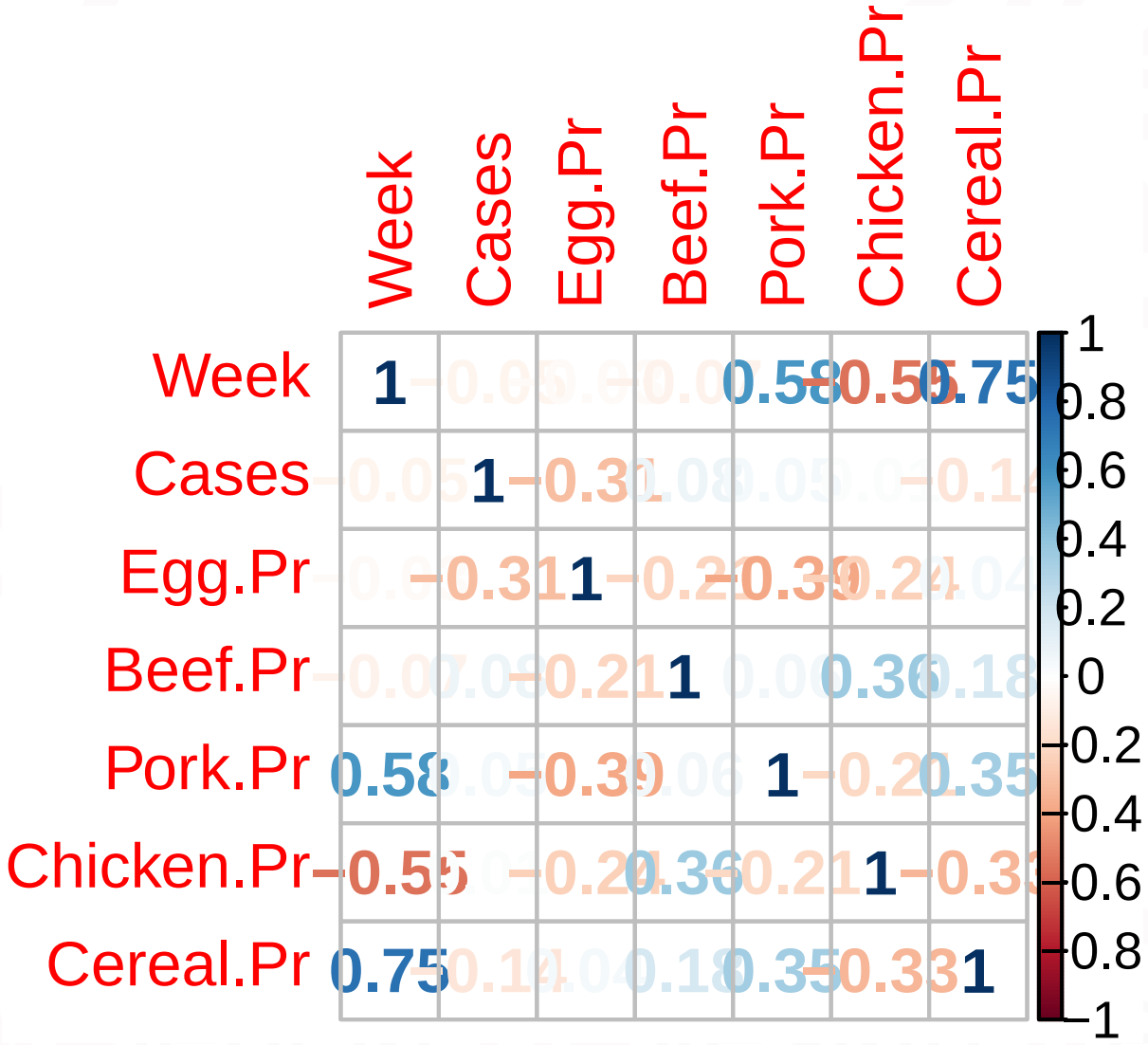
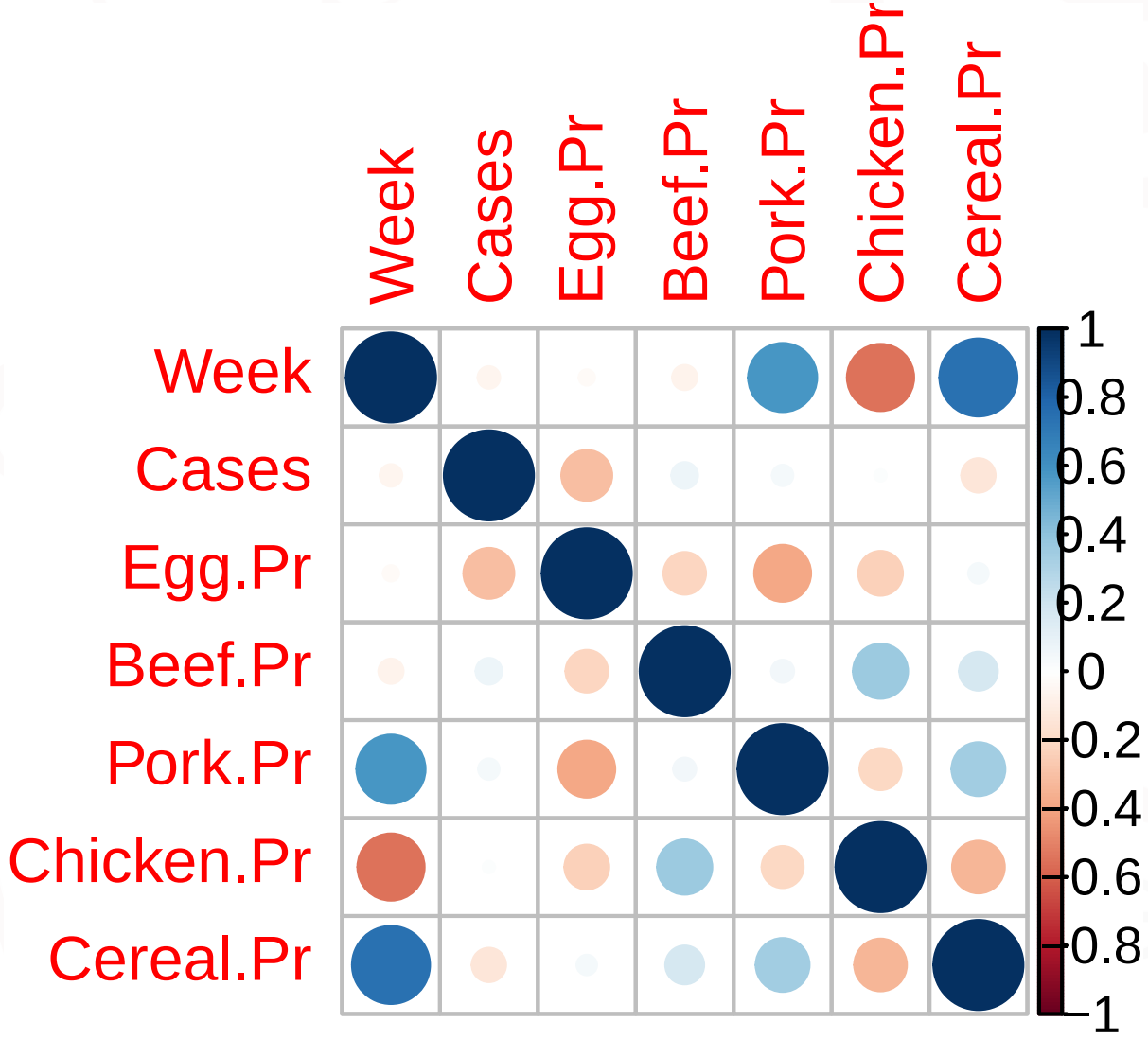

Week
Cases
Egg.Pr
Beef.Pr
Pork.Pr
Chicken.Pr
Cereal.Pr
1
0.58
0.75
1
0.31
1
0.21
1
0.36
1
1
1
1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Table 1: Correlaciones

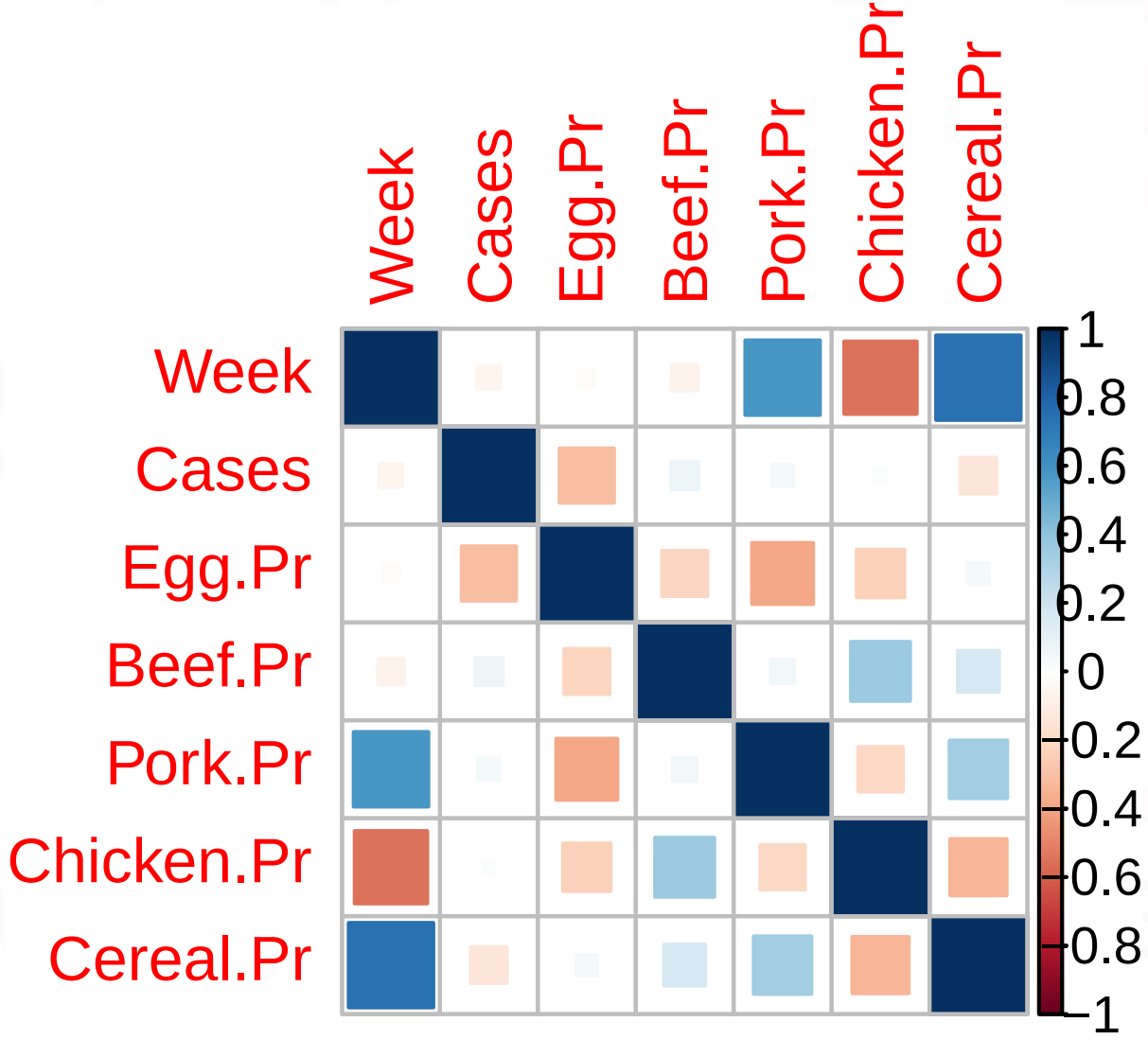
	Week	Cases	Egg.Pr	Beef.Pr	Pork.Pr	Chicken.Pr	Cereal.Pr
Week	1.00	-0.05	-0.03	-0.07	0.58	-0.55	0.75
Cases	-0.05	1.00	-0.31	0.08	0.05	0.01	-0.14
Egg.Pr	-0.03	-0.31	1.00	-0.21	-0.39	-0.24	0.04
Beef.Pr	-0.07	0.08	-0.21	1.00	0.06	0.36	0.18
Pork.Pr	0.58	0.05	-0.39	0.06	1.00	-0.21	0.35
Chicken.Pr	-0.55	0.01	-0.24	0.36	-0.21	1.00	-0.33
Cereal.Pr	0.75	-0.14	0.04	0.18	0.35	-0.33	1.00


Week
Cases
Egg.Pr
Beef.Pr
Pork.Pr
Chicken.Pr
Cereal.Pr
1
0.58
0.75
1
0.31
1
0.21
1
0.36
1
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1

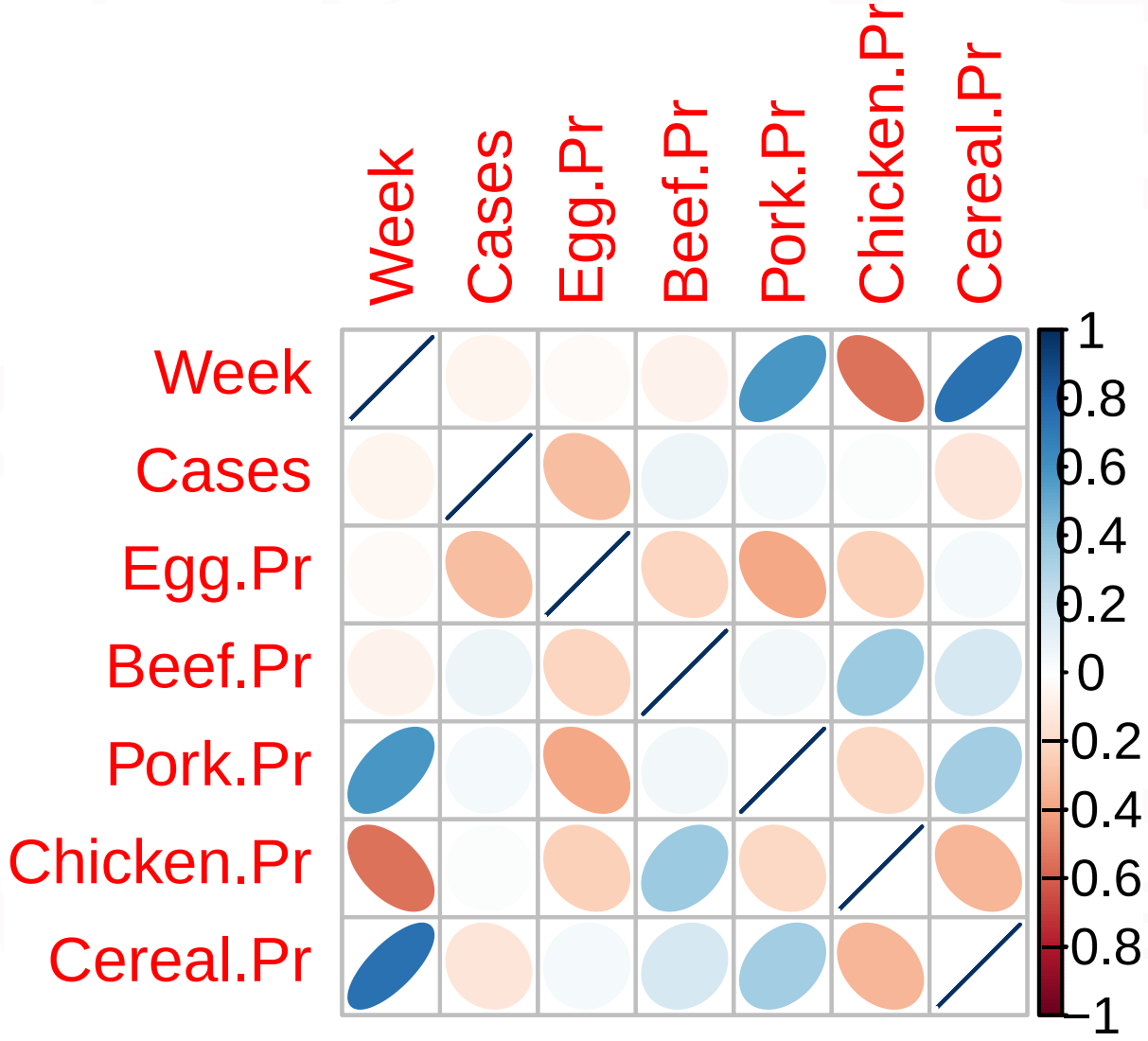
```
corrplot(M, method = "circle")
```



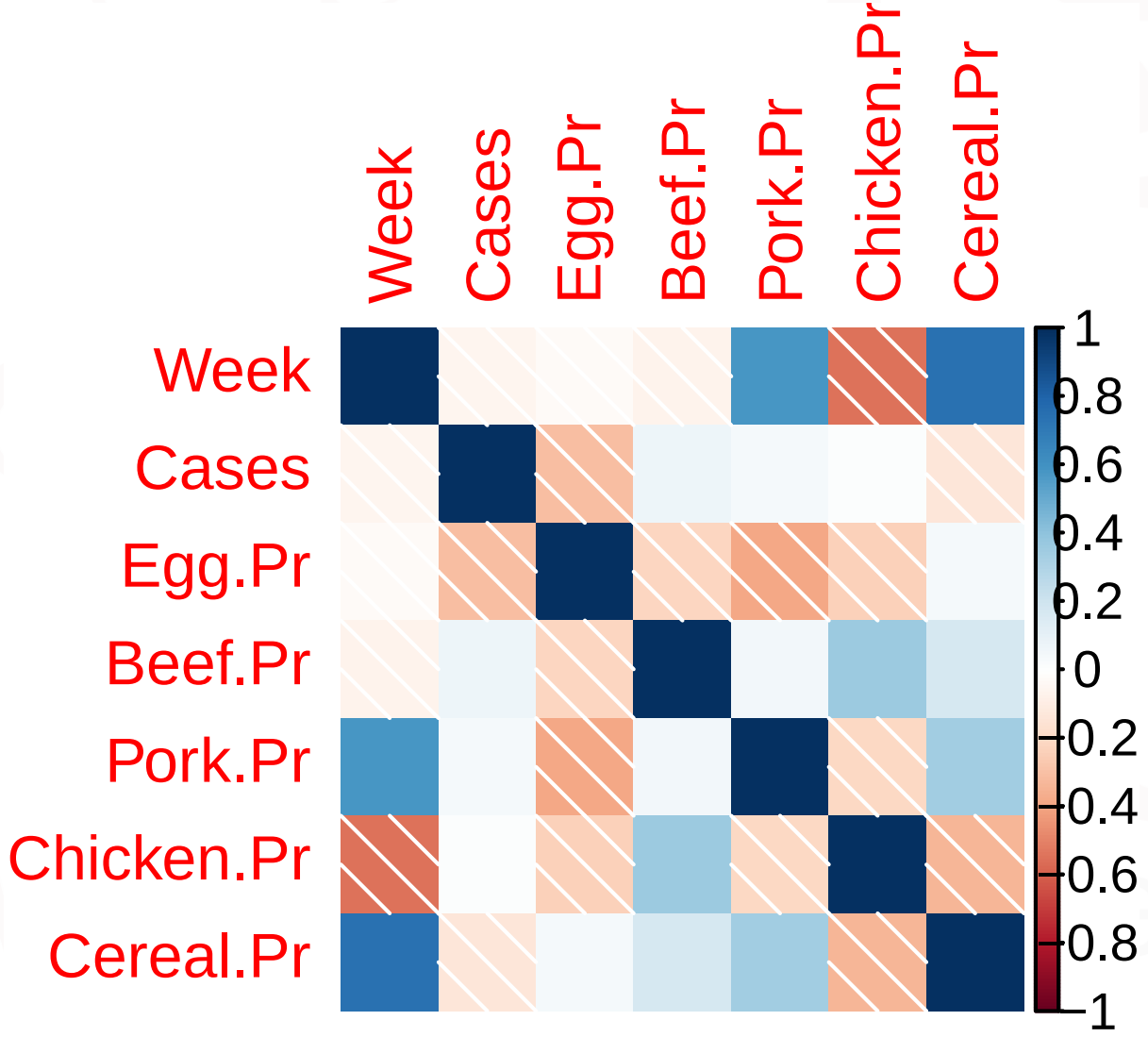
```
corrplot(M, method = "square")
```



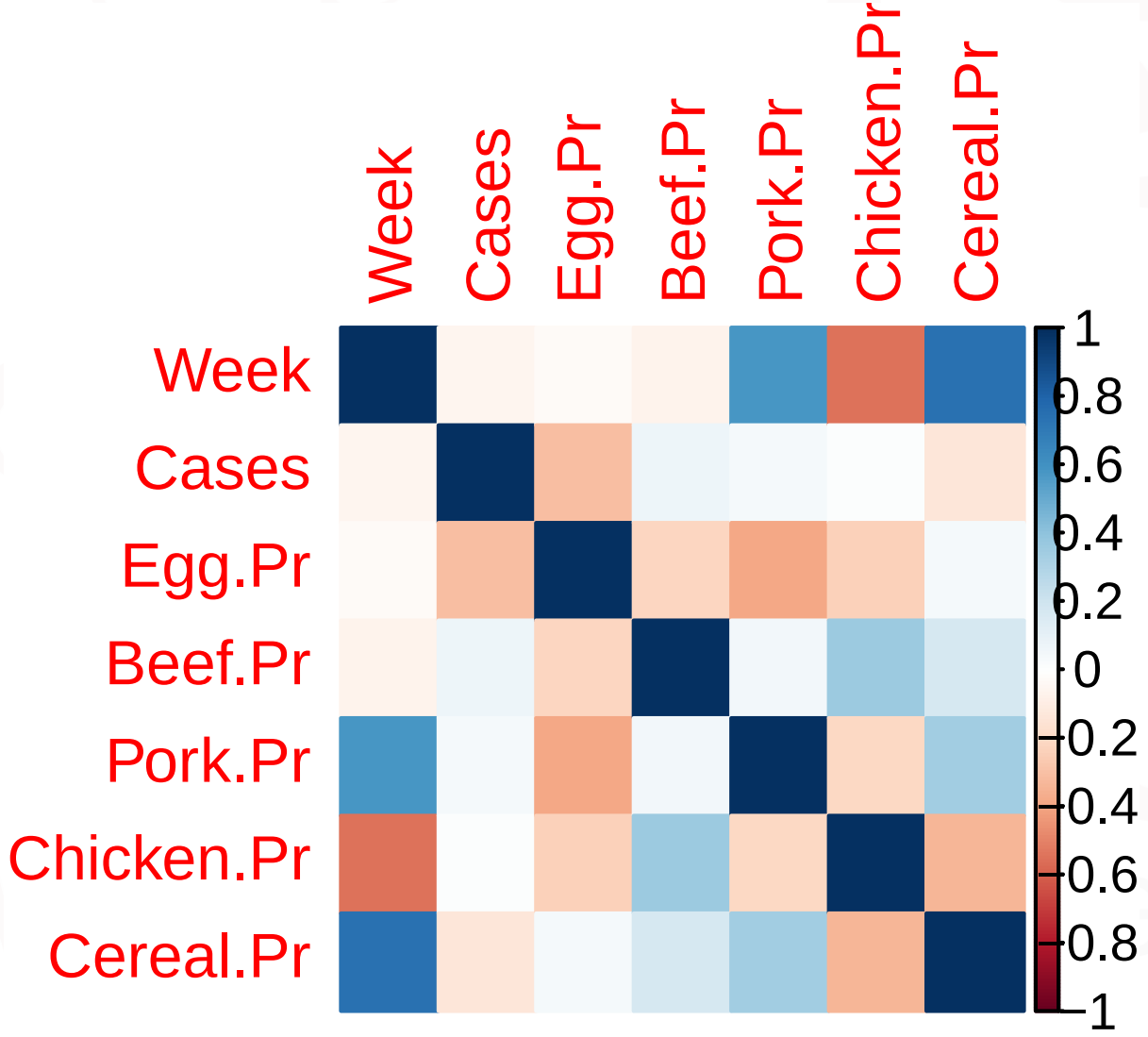
```
corrplot(M, method = "ellipse")
```



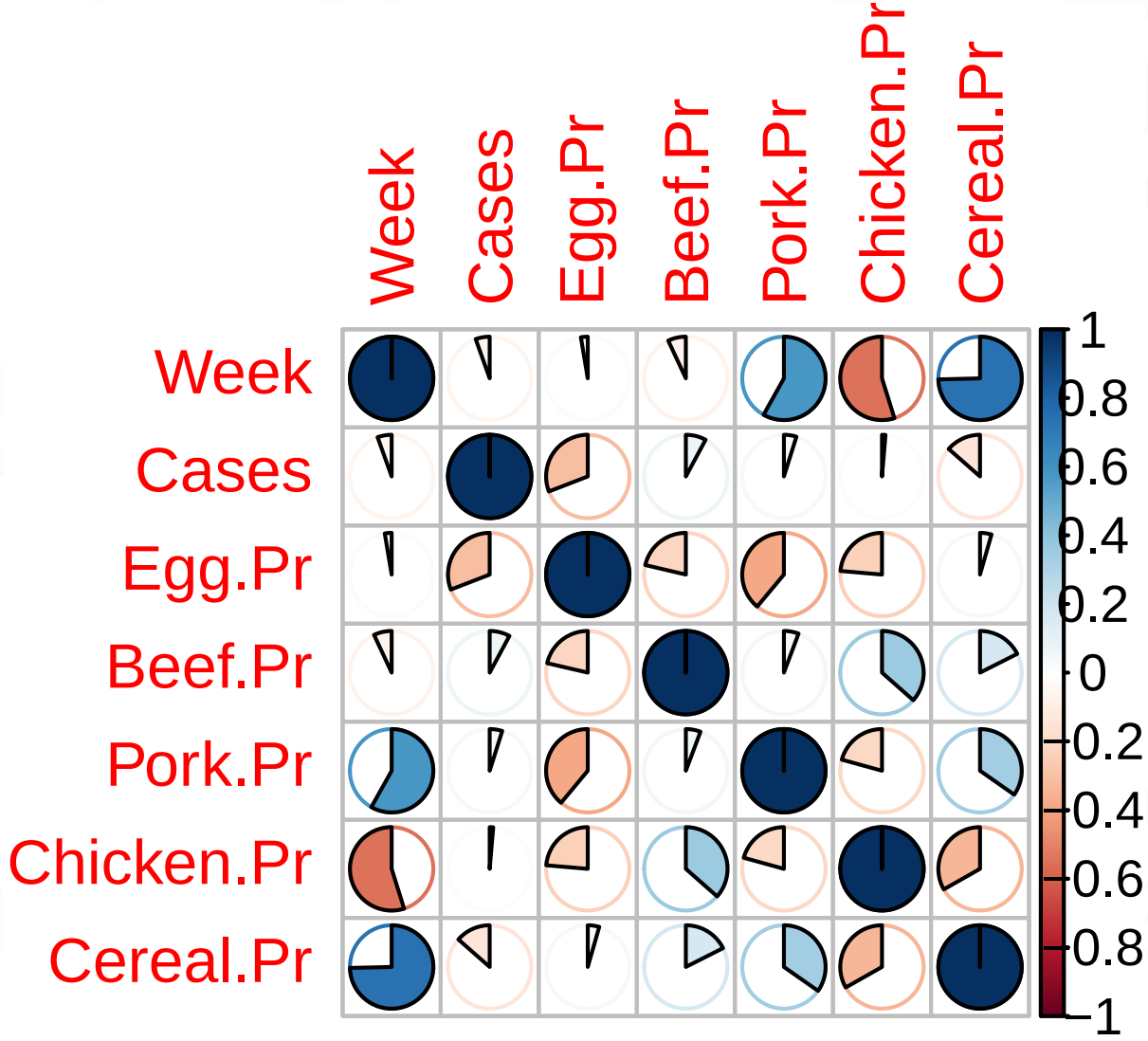
```
corrplot(M, method = "shade")
```



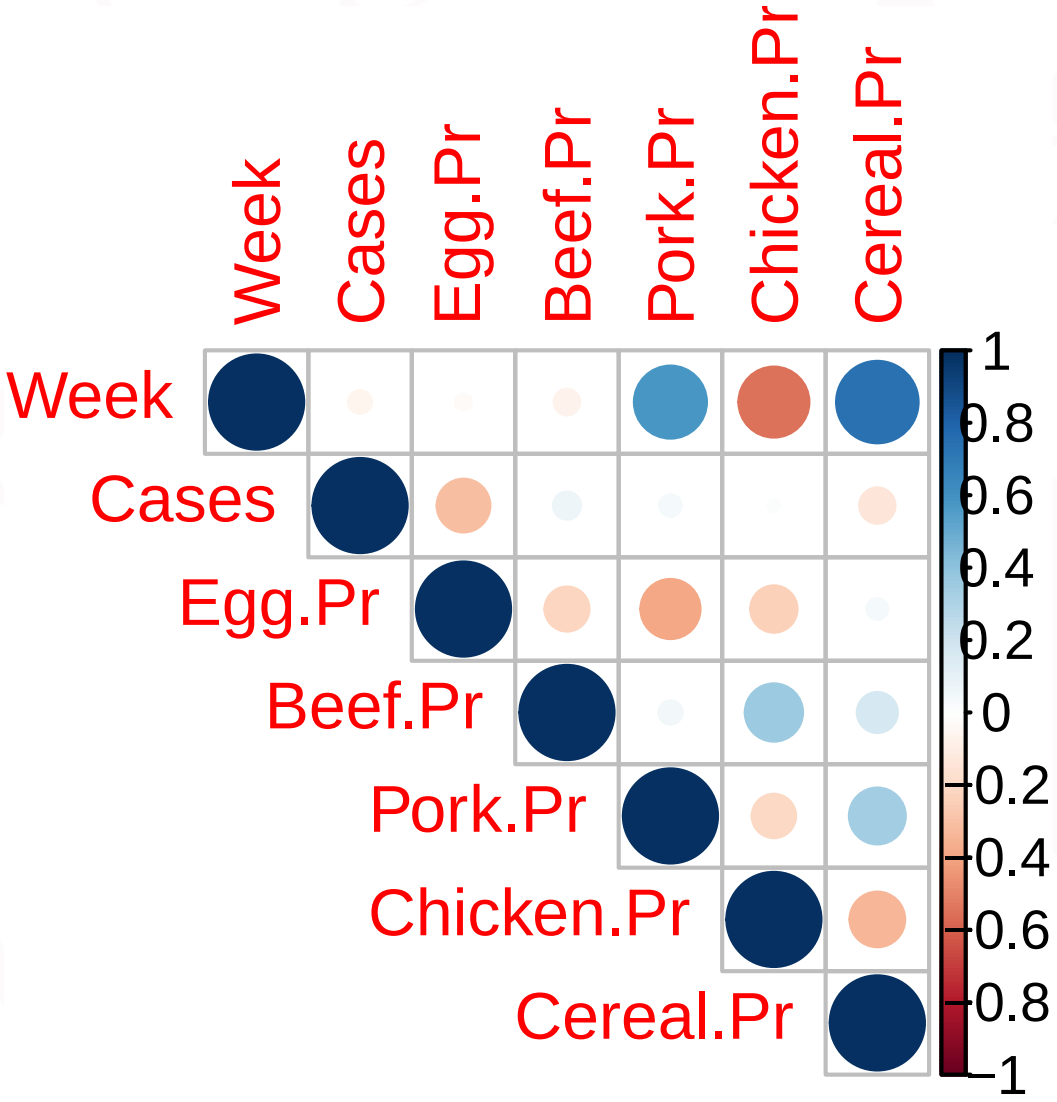
```
corrplot(M, method = "color")
```



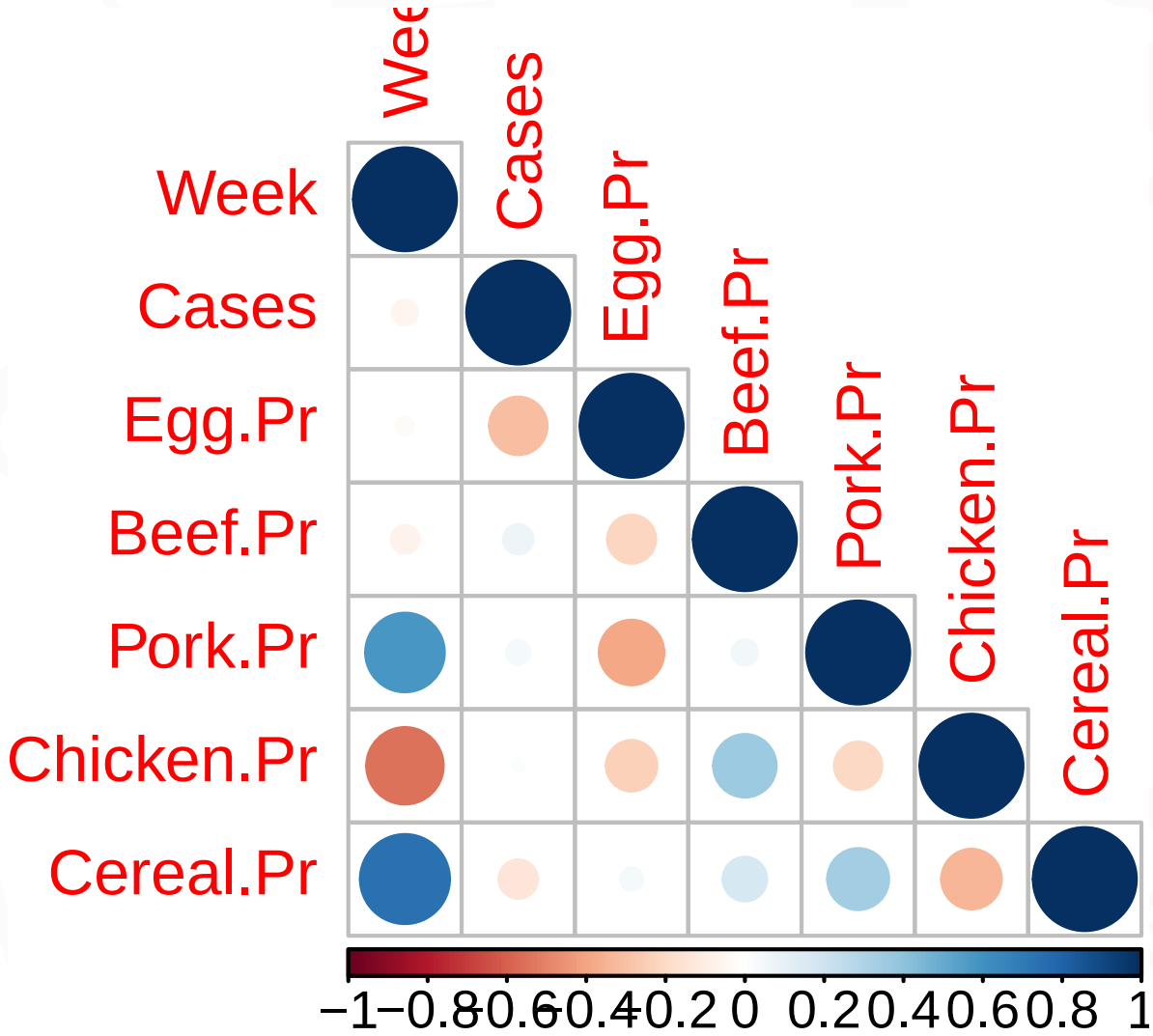
```
corrplot(M, method = "pie")
```



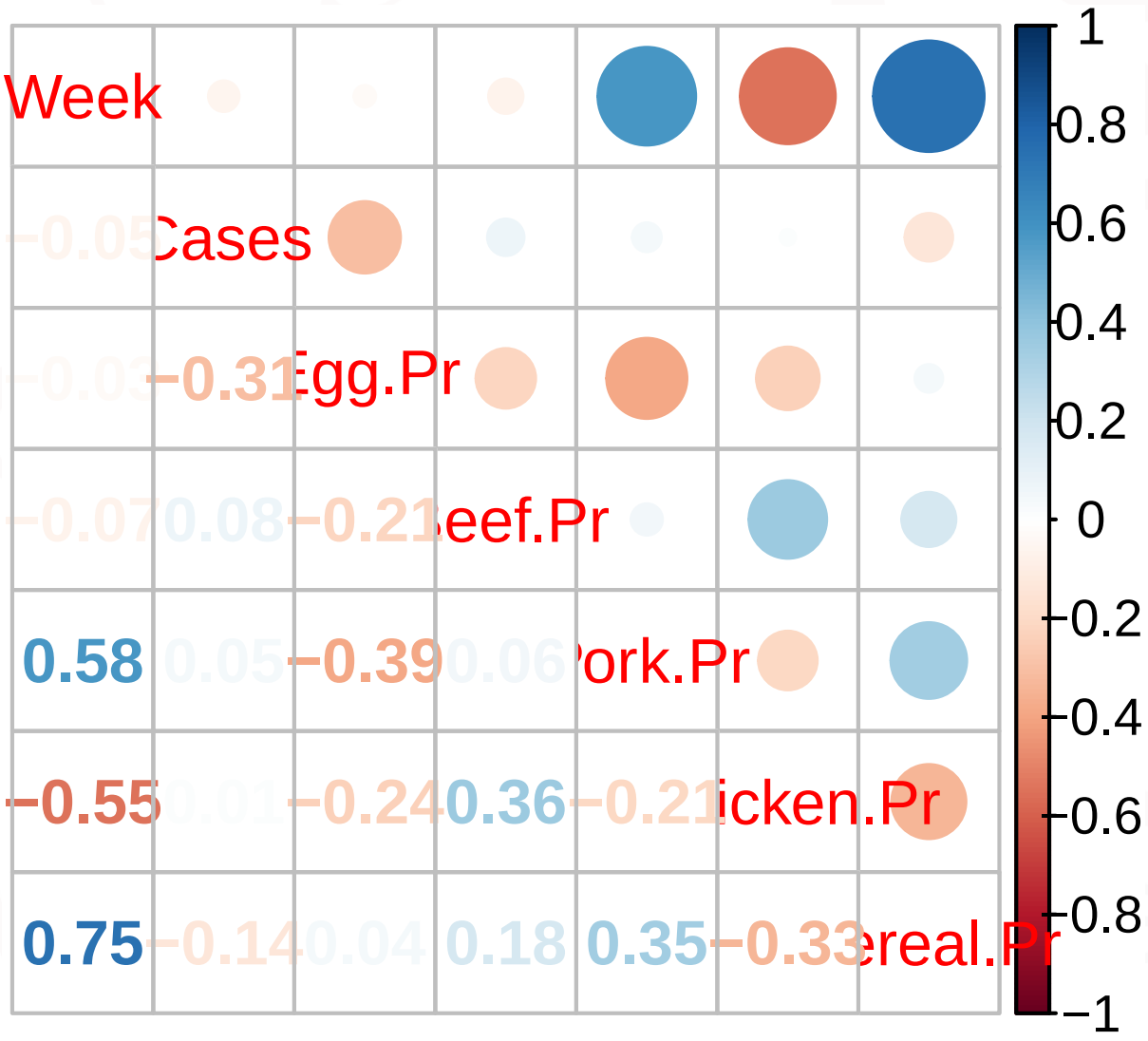
```
corrplot(M, type = "upper")
```



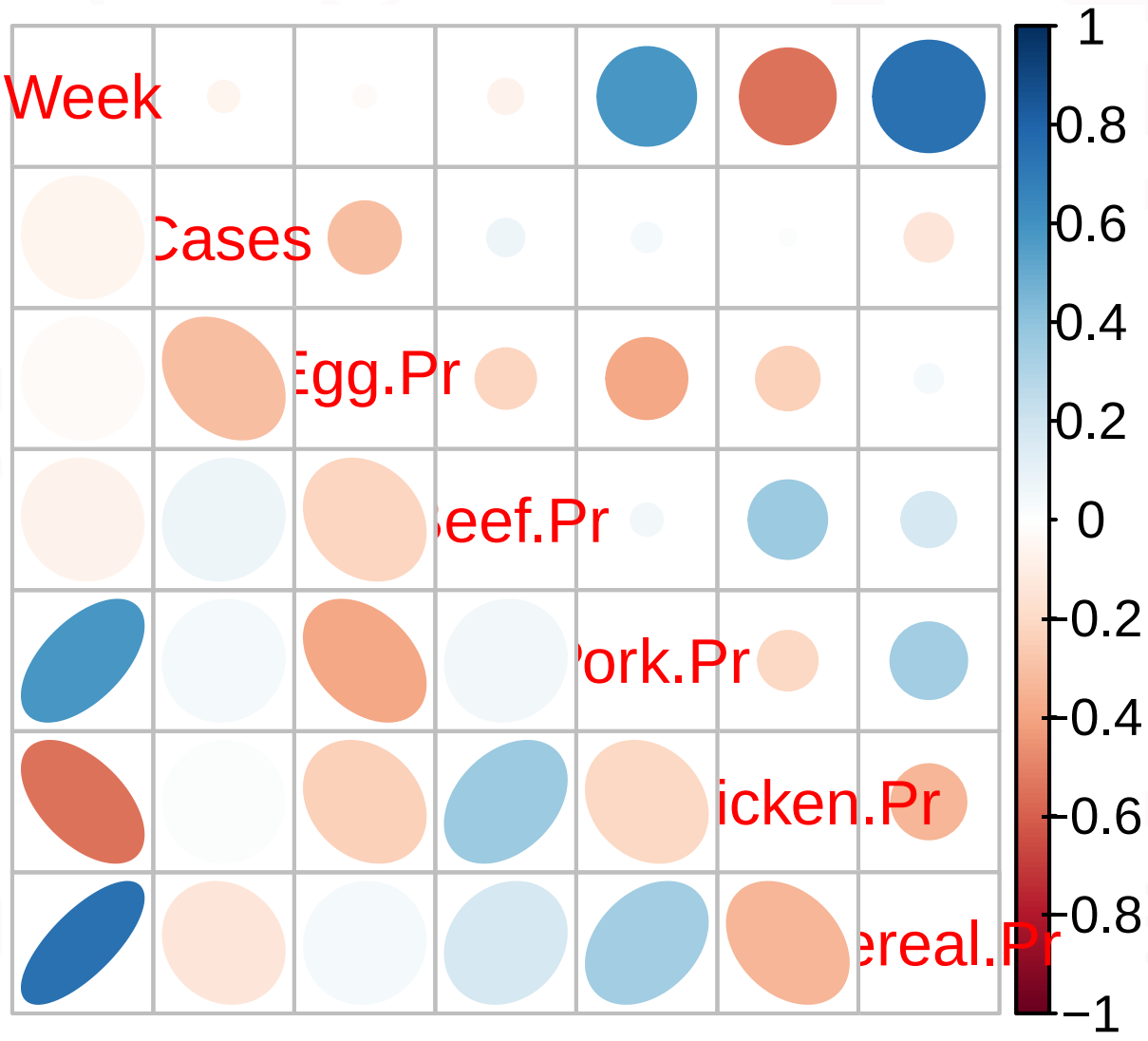
```
corrplot(M, type = "lower")
```



```
corrplot.mixed(M)
```



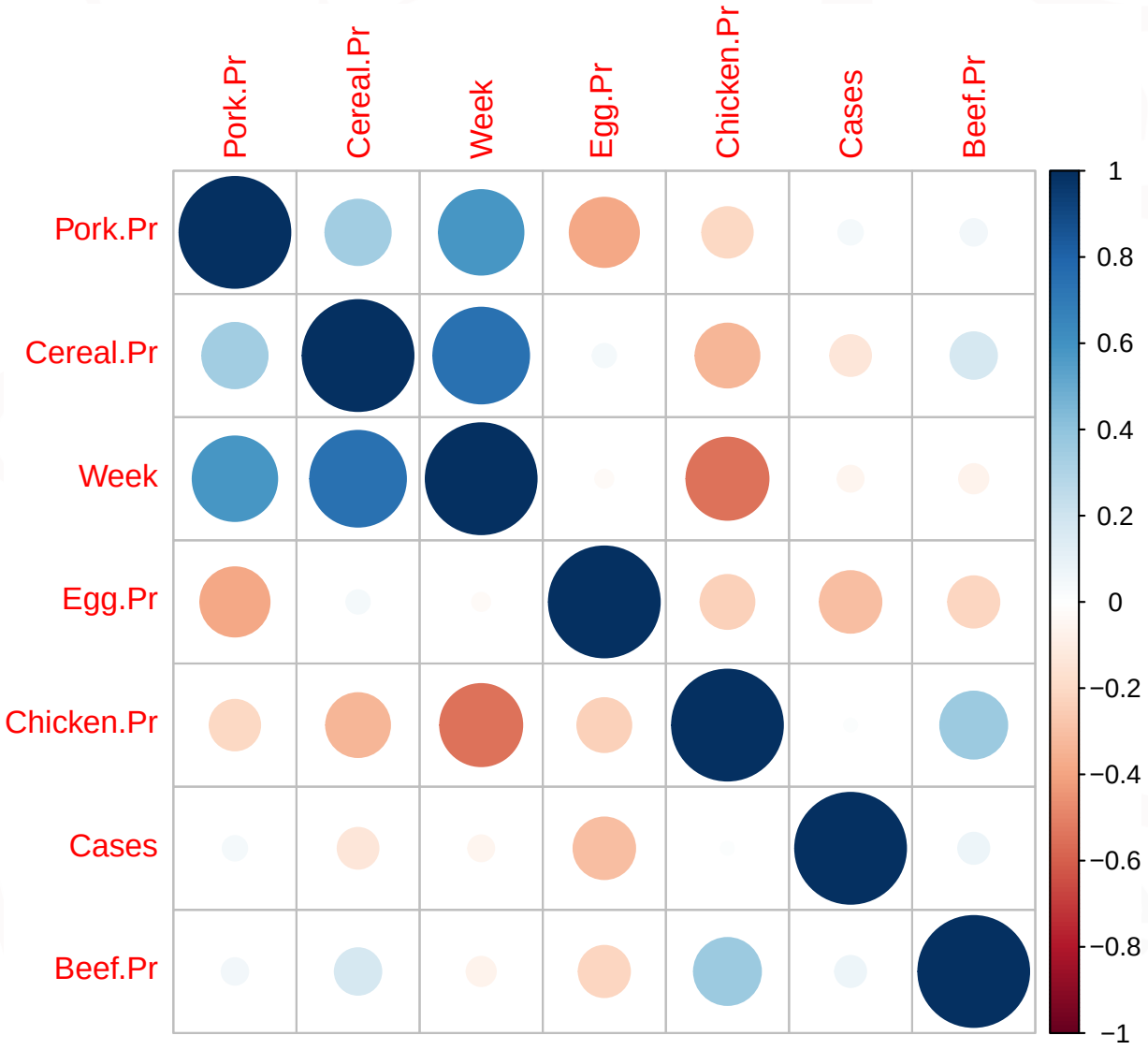
```
corrplot.mixed(M, lower = "ellipse", upper = "circle")
```



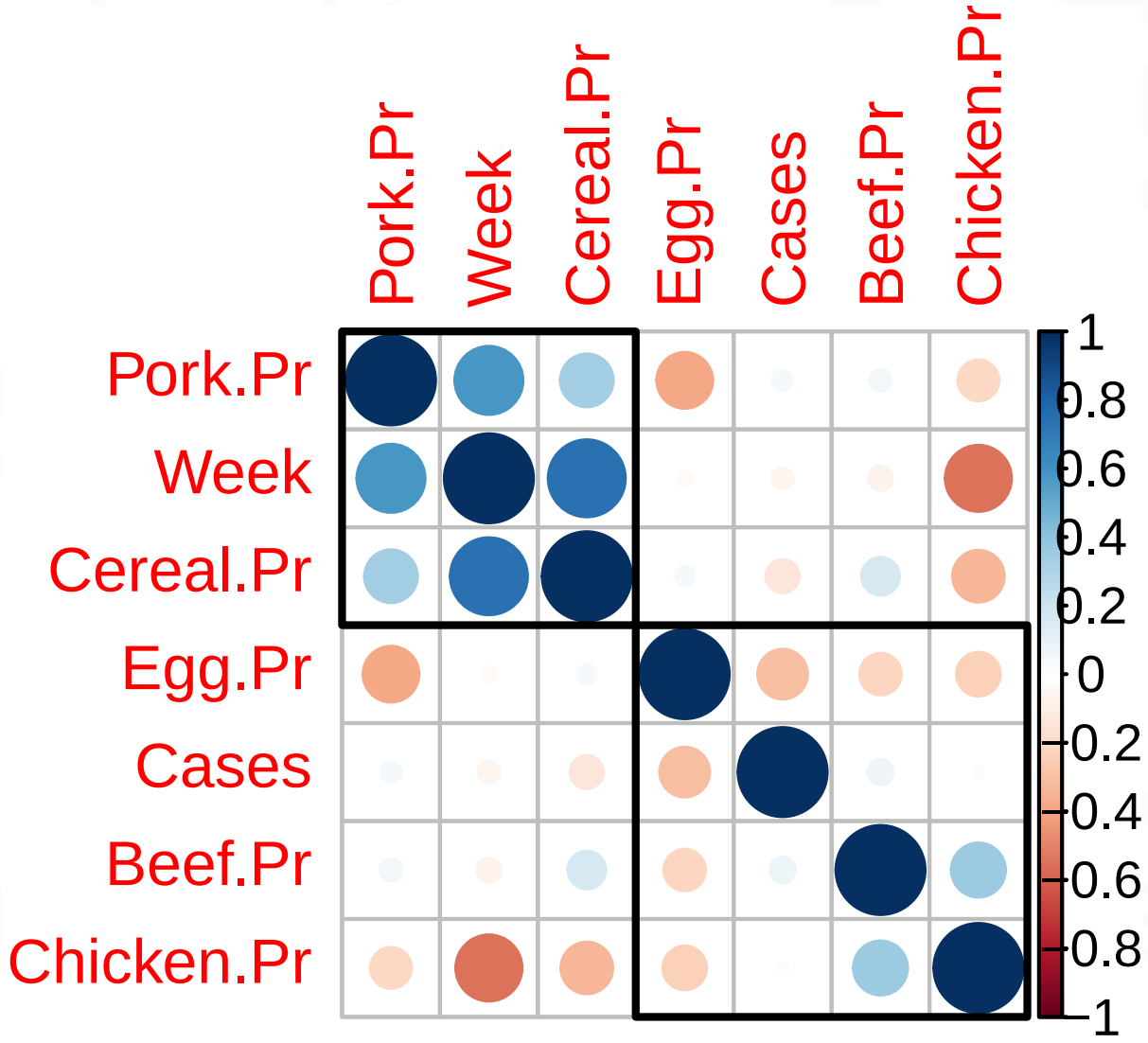
Reordenado la correlación

- **FPC** for the first principal component order.
- **hclust** for hierarchical clustering order, and “hclust.method” for the agglomeration method to be used . “hclust.method” should be one of “ward”, “single”, “complete”, “average”,
- **mcquitty**, “median” or “centroid”.
- **alphabet** for alphabetical order.

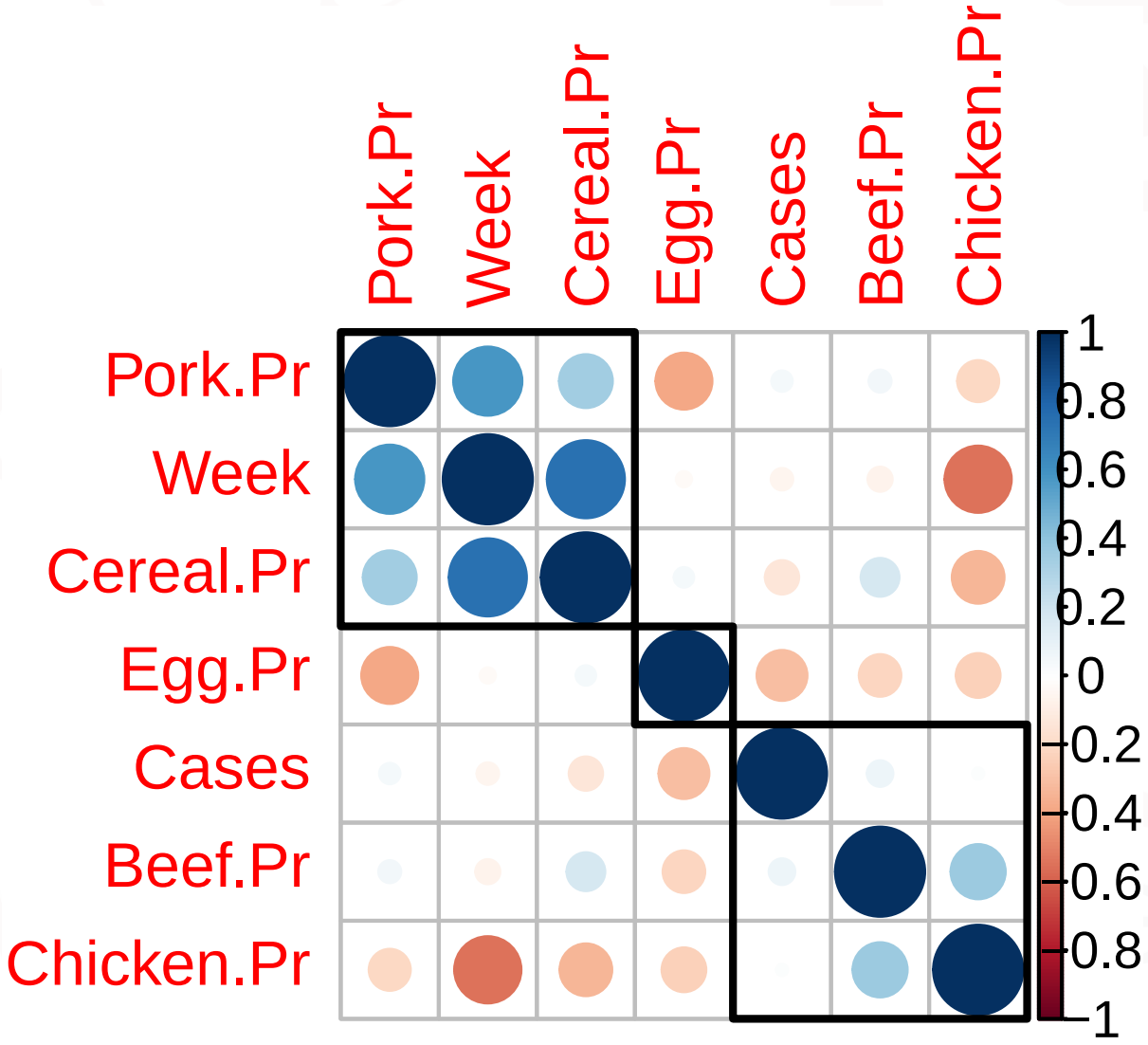
```
corrplot(M, order = "AOE")
```



```
corrplot(M, order = "hclust", addrect = 2)
```

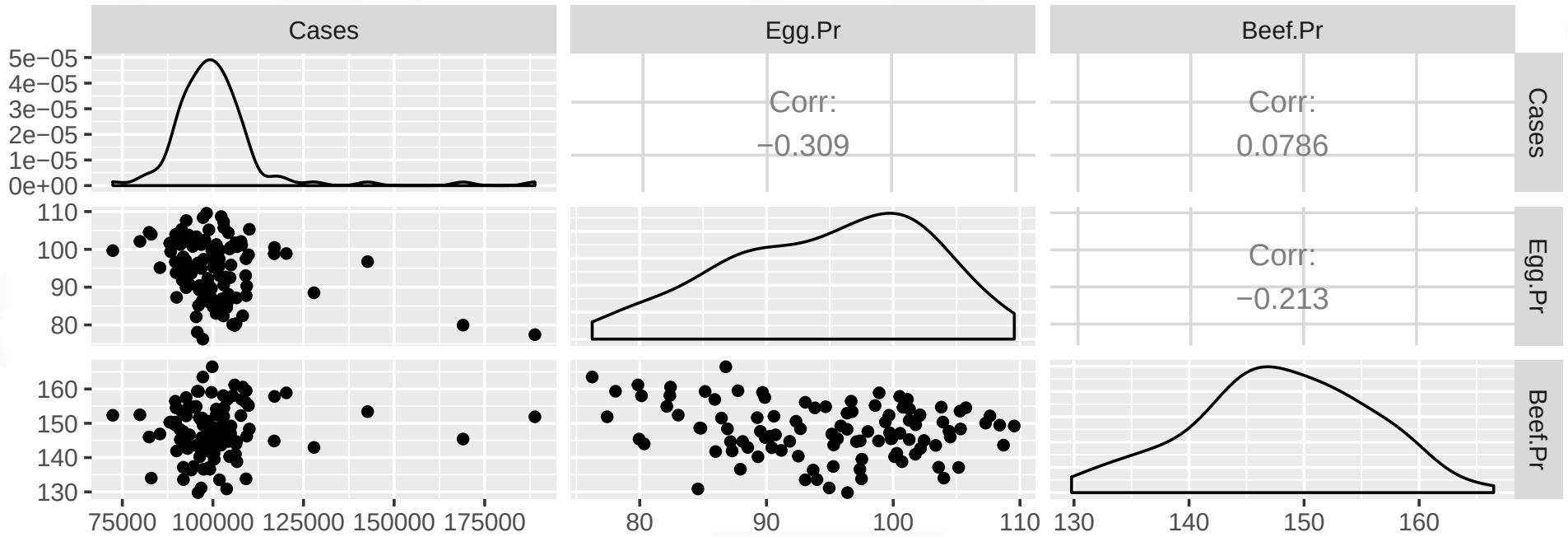


```
corrplot(M, order = "hclust", addrect = 3)
```



ggparis

```
library(GGally)
ggpairs(Eggs[, 4:10], color = Easter)
```


Cases
Egg.Pr
Beef.Pr
Corr:
−0.309
Corr:
0.0786
Corr:
−0.213
Cases
Egg.Pr
Beef.Pr
5e−05
4e−05
3e−05
2e−05
1e−05
0e+00
110
100
90
80
160
150
140
130
75000
100000
125000
150000
175000
80
90
100
110
130
140
150
160

```
p_ <- GGally::print_if_interactive
pm <- ggpairs(Eggs, columns = 5:7, ggplot2::aes(colour = Easter))
p_(pm)
ggsave("img/ppm.pdf")
```

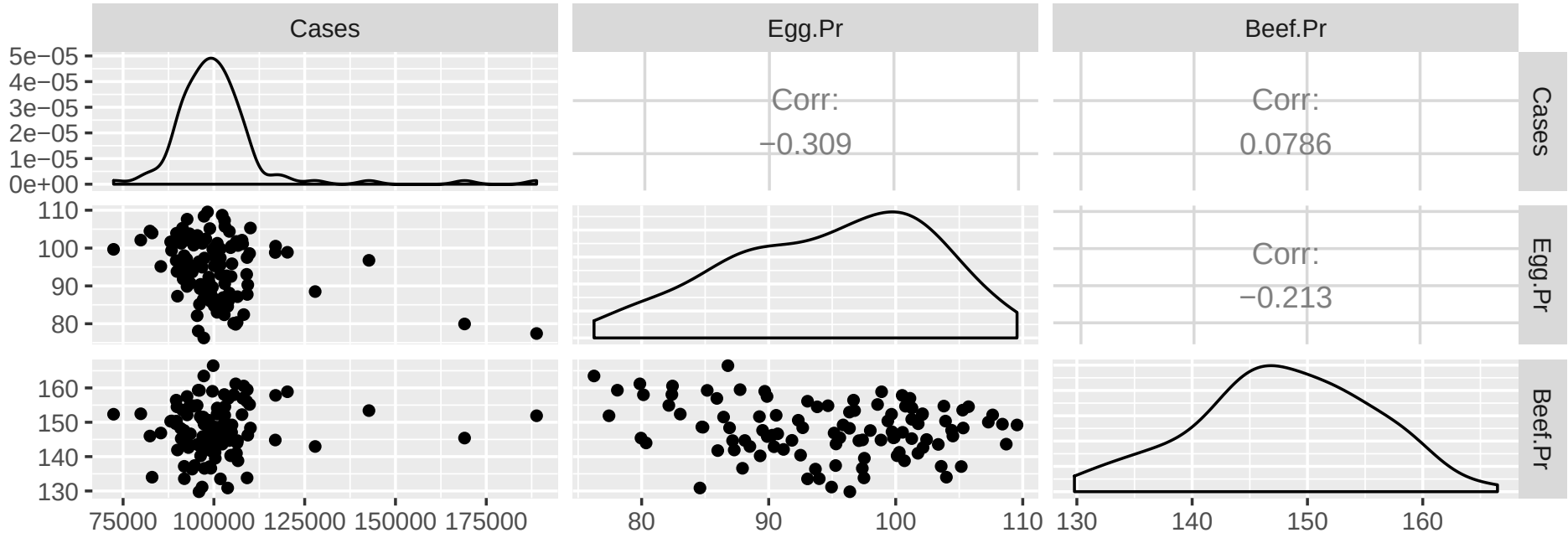
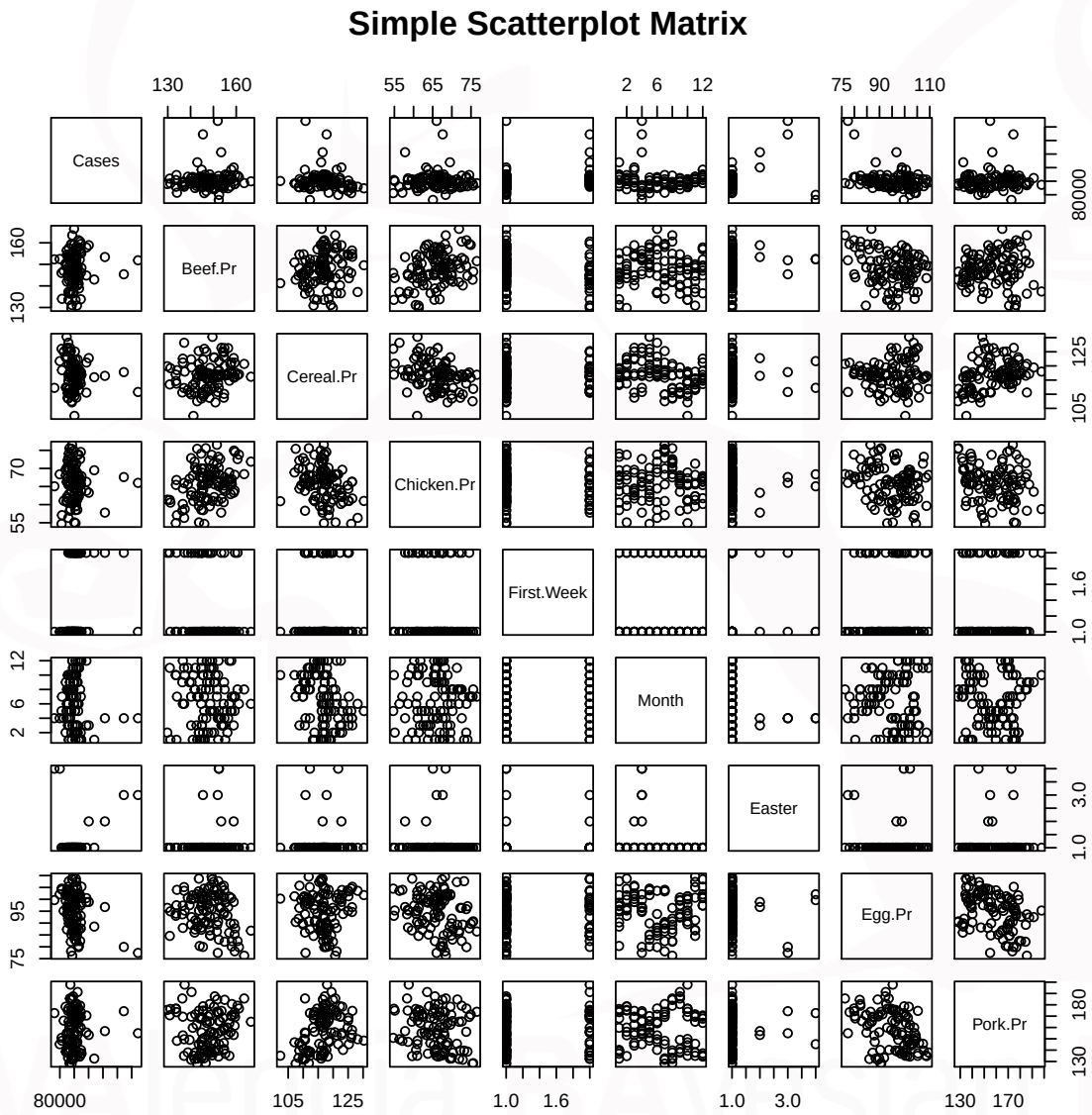
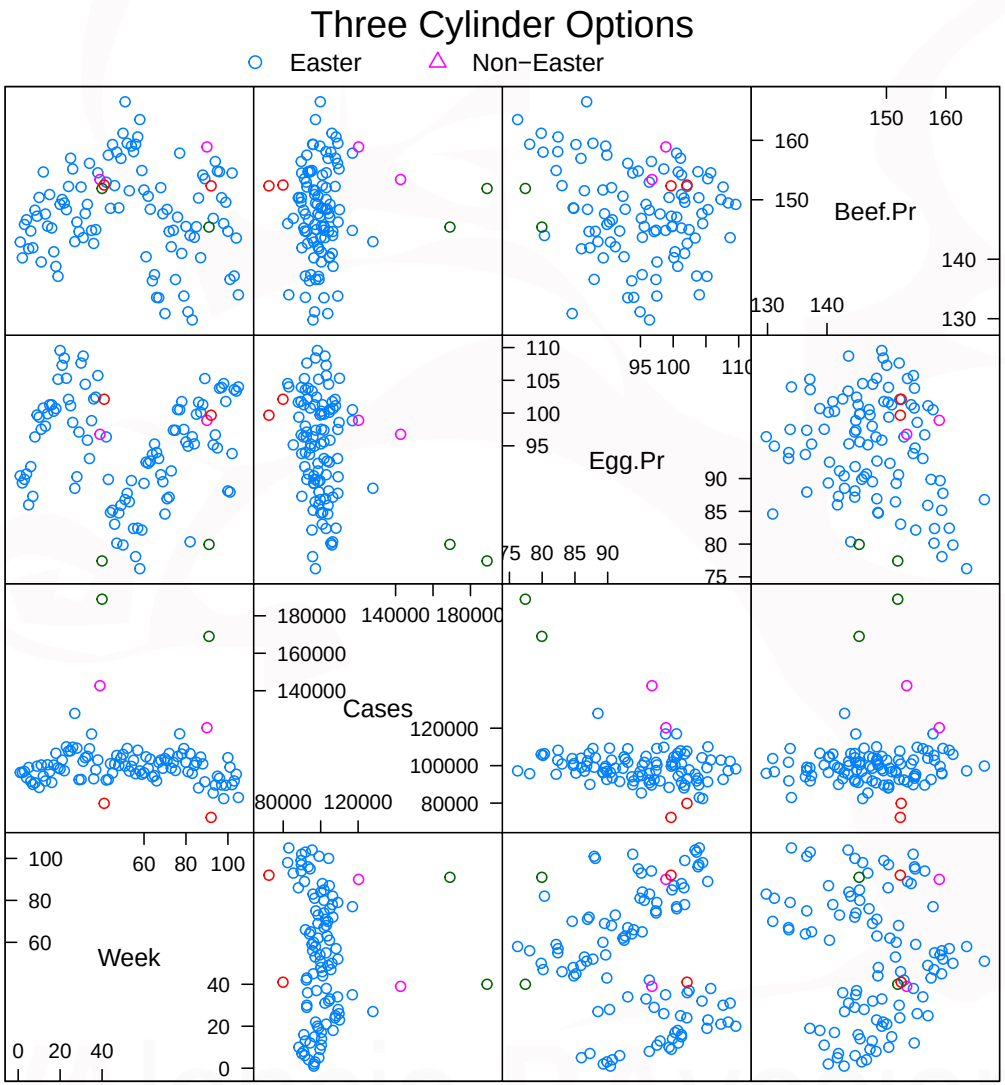

Cases
Egg.Pr
Beef.Pr
Corr:
−0.309
Corr:
0.0786
Corr:
−0.213
Cases
Egg.Pr
Beef.Pr
5e−05
4e−05
3e−05
2e−05
1e−05
0e+00
110
100
90
80
160
150
140
130
75000
100000
125000
150000
175000
80
90
100
110
130
140
150
160

Figure 2

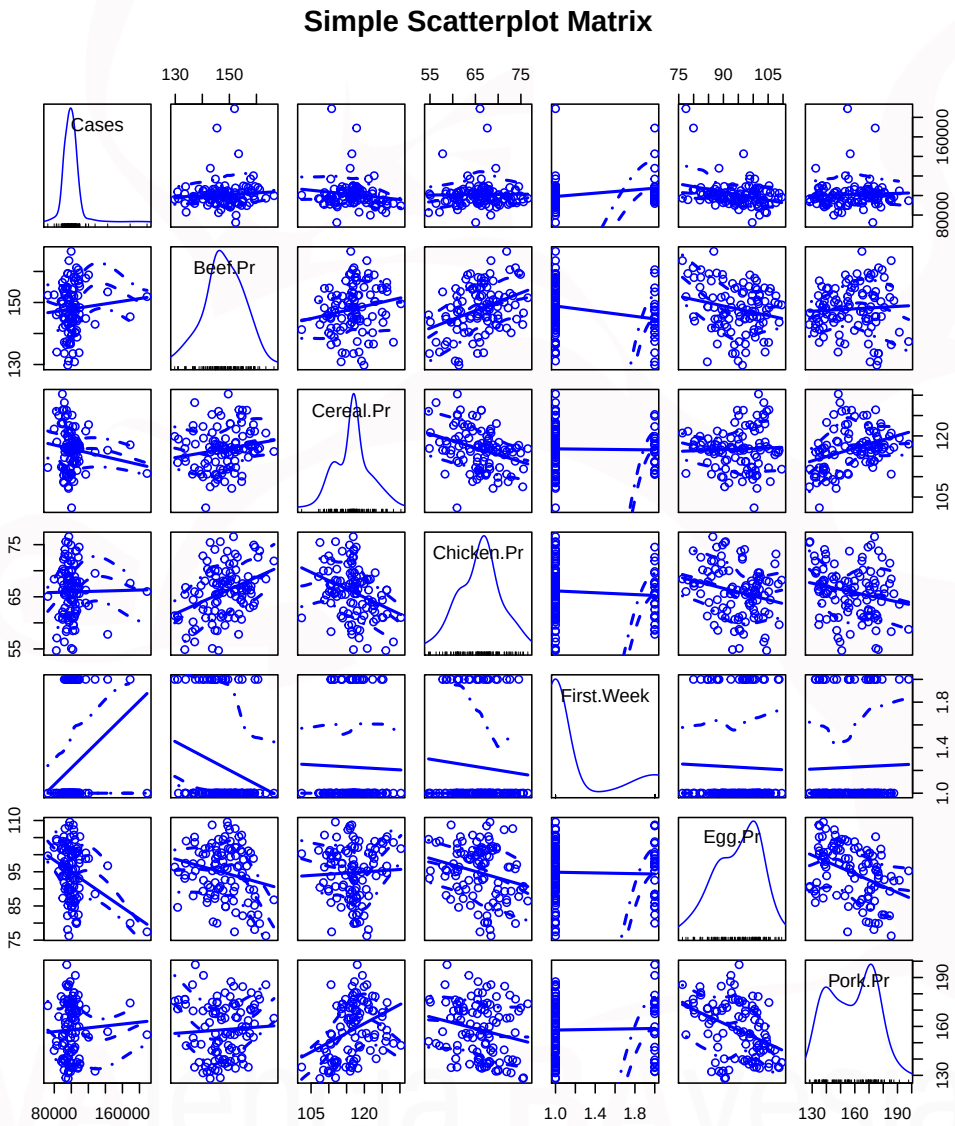
```
pairs(~Cases +Beef.Pr + Cereal.Pr +
        Chicken.Pr + First.Week +
        Month + Easter + Egg.Pr+Pork.Pr,
      data=Eggs,   main="Simple Scatterplot Matrix")
```


Simple Scatterplot Matrix
Cases
Beef.Pr
Cereal.Pr
Chicken.Pr
First.Week
Month
Easter
Egg.Pr
Pork.Pr

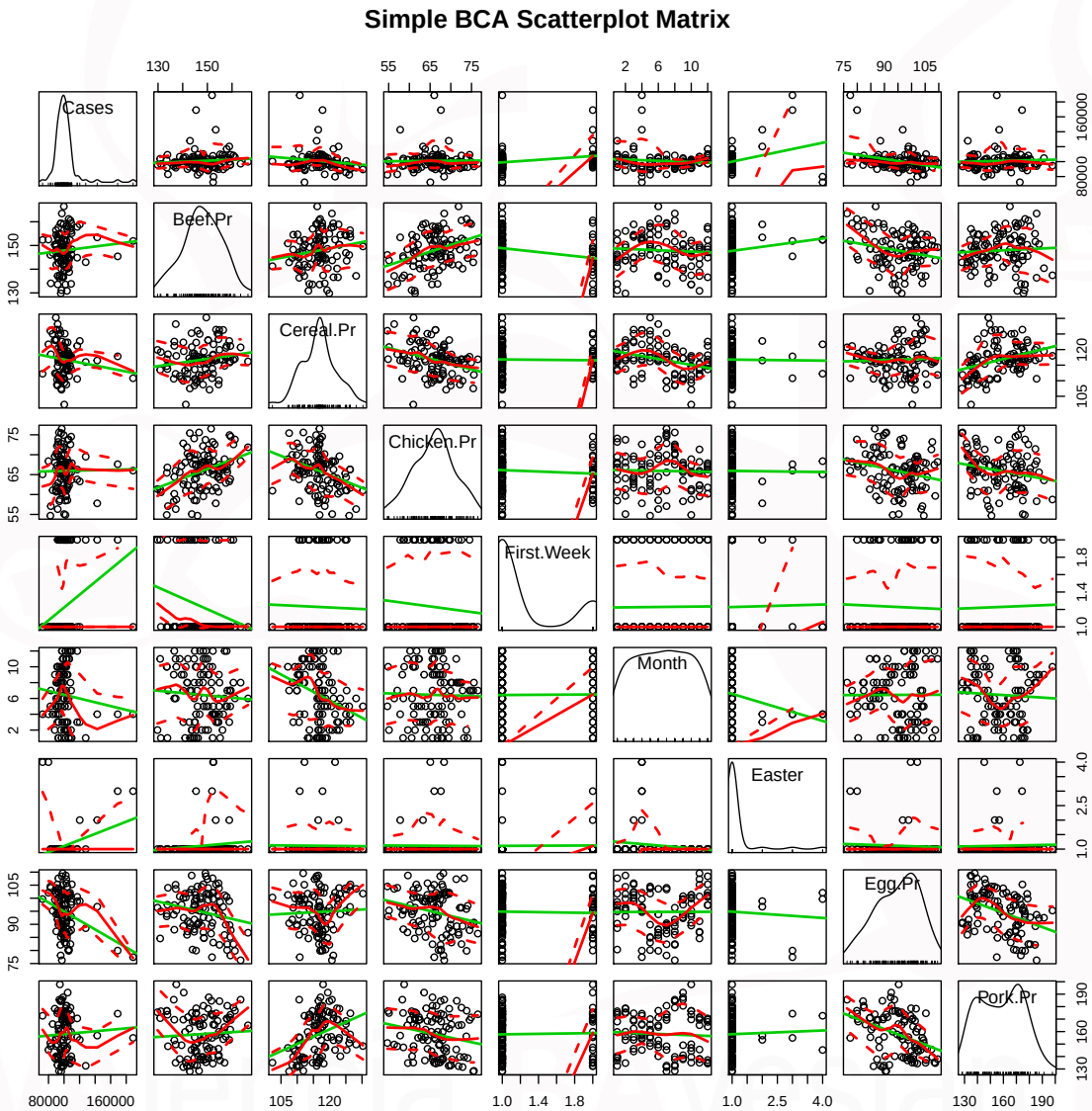
```
library(lattice)
super.sym <- trellis.par.get("superpose.symbol")
super.sym$pch <- 1:length(super.sym$pch) # change this
                                         # to specify symbols
splom(Eggs[c(1,5,6,7)], groups=Eggs$Easter, data=Eggs,
    panel=panel.superpose,
   key=list(title="Three Cylinder Options",
   columns=3,
   points=list(pch=super.sym$pch[1:2],
   col=super.sym$col[1:2]),
   text=list(c("Easter","Non-Easter"))))
```


Three Cylinder Options
Easter
Non−Easter
Beef.Pr
Egg.Pr
Cases
Week
Scatter Plot Matrix

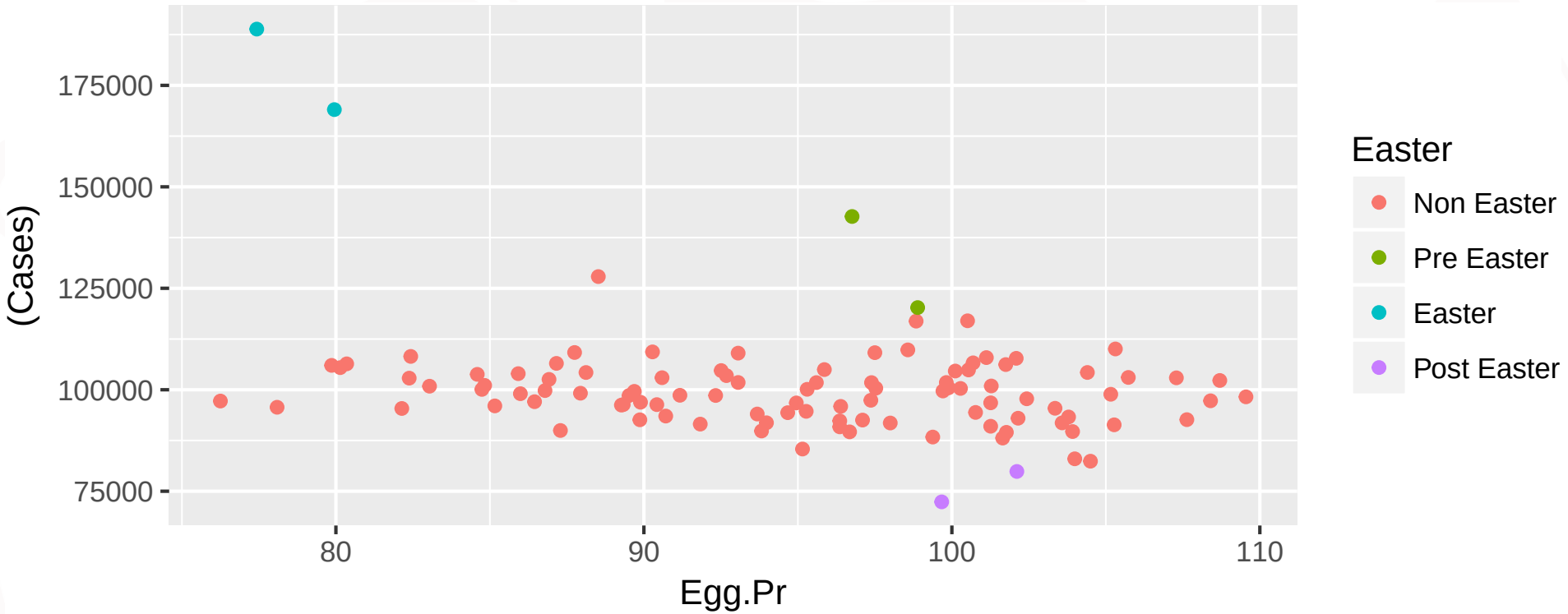
```
library(car)
scatterplotMatrix(~Cases+Beef.Pr
                  + Cereal.Pr +    Chicken.Pr +
                    First.Week +
                     Egg.Pr+ Pork.Pr,
                  data=Eggs,
                main="Simple Scatterplot Matrix")
```


Simple Scatterplot Matrix
Cases
Beef.Pr
Cereal.Pr
Chicken.Pr
First.Week
Egg.Pr
Pork.Pr

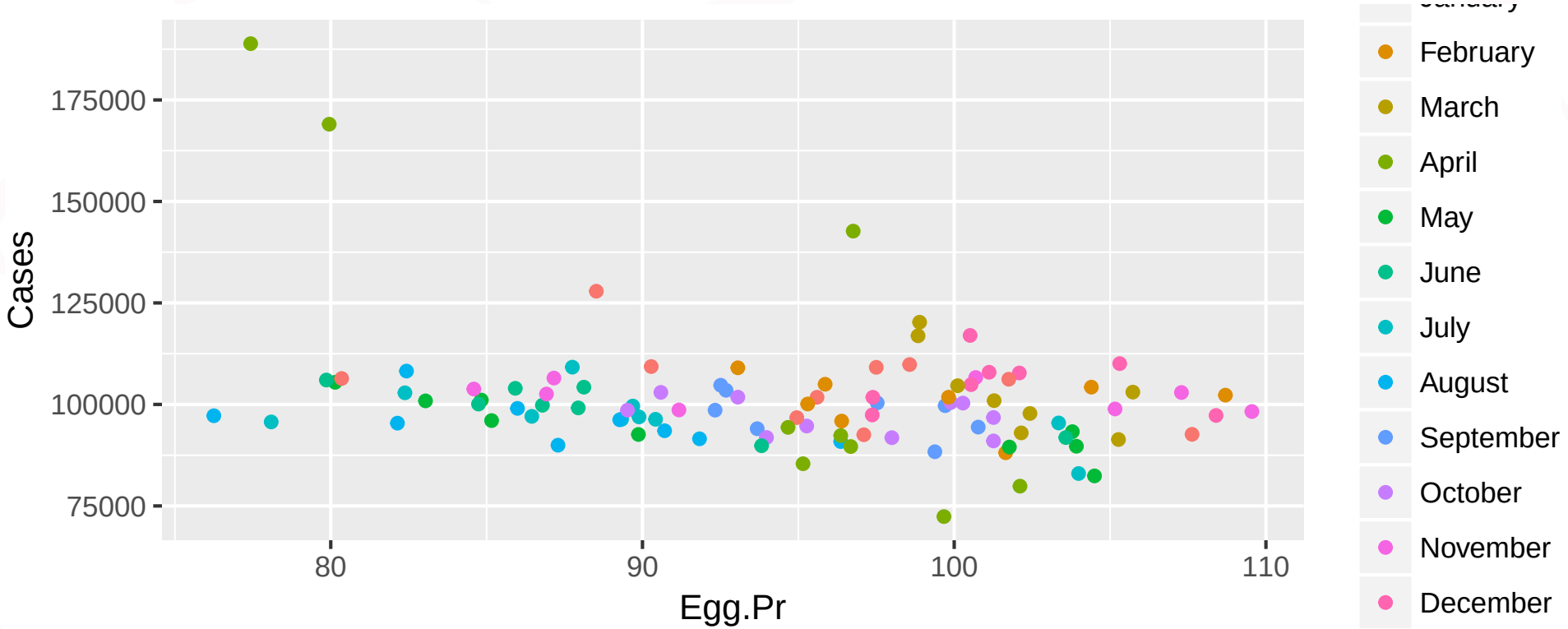
```
library(BCA)
scatterplotMatrixBCA(~Cases      +Beef.Pr
                     + Cereal.Pr +  Chicken.Pr +
                       First.Week +      Month + Easter +
                       Egg.Pr+  Pork.Pr, data=Eggs,
                   main="Simple BCA Scatterplot Matrix")
```


Simple BCA Scatterplot Matrix
Cases
Beef.Pr
Cereal.Pr
Chicken.Pr
First.Week
Month
Easter
Egg.Pr
Pork.Pr

```
library(ggplot2)
ggplot(Eggs, aes(y = (Cases), x = Egg.Pr, color = Easter)) +
    geom_point()
```

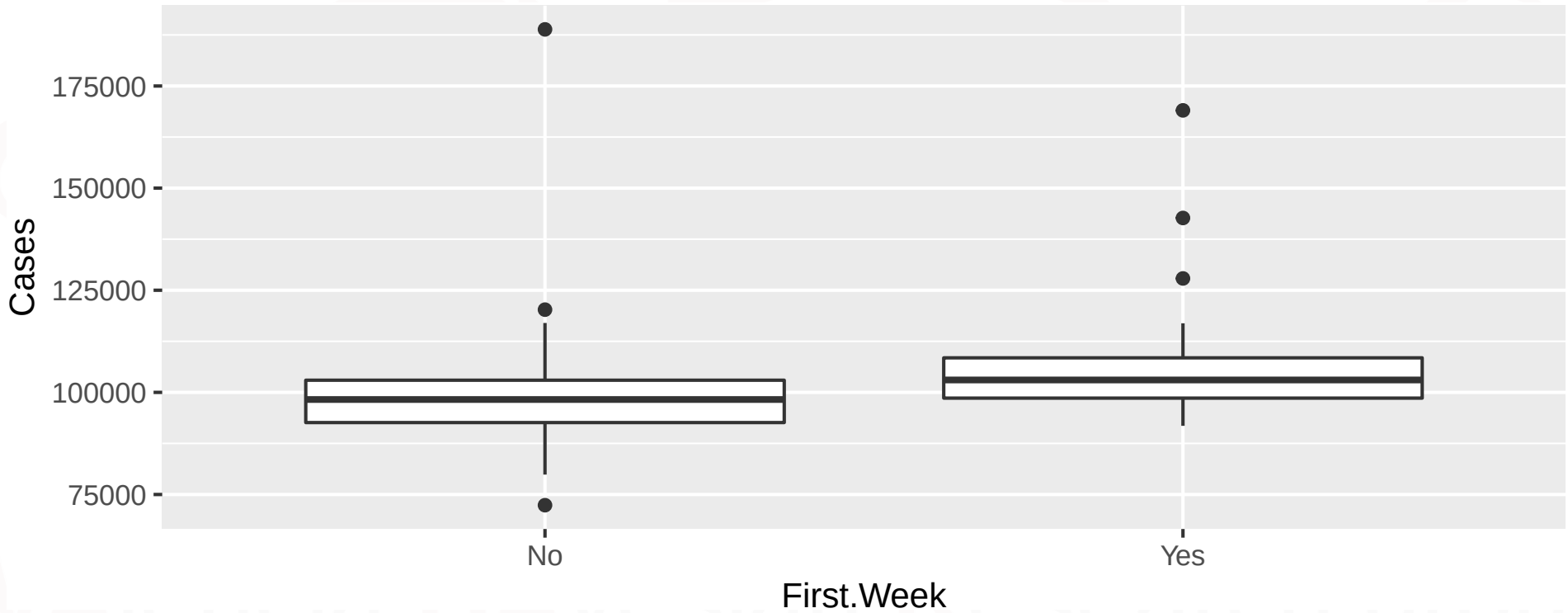


```
library(ggplot2)
ggplot(Eggs, aes(y = Cases, x = Egg.Pr, color = Month)) + geom_point
```

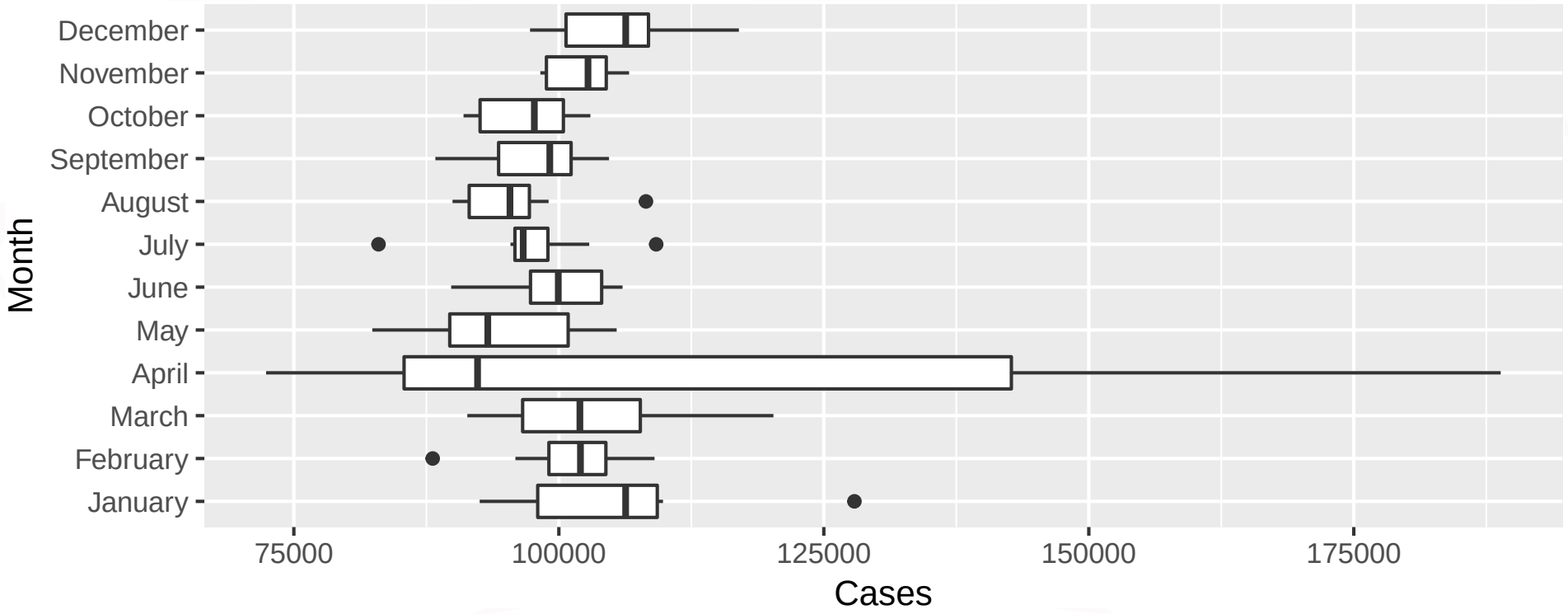


BoxPlots

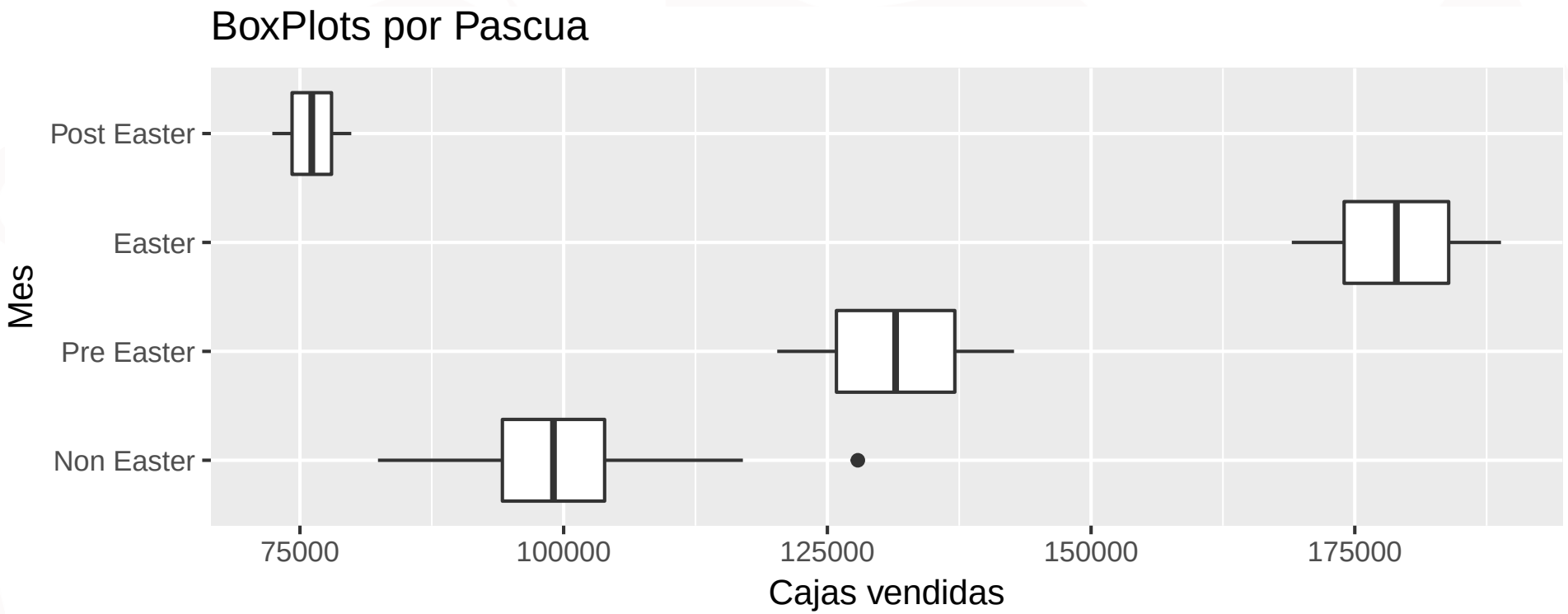
```
ggplot(Eggs, aes(y = Cases, x = First.Week)) + geom_boxplot()
```



```
ggplot(Eggs, aes(y = Cases, x = Month)) + geom_boxplot() + coord_fli
```

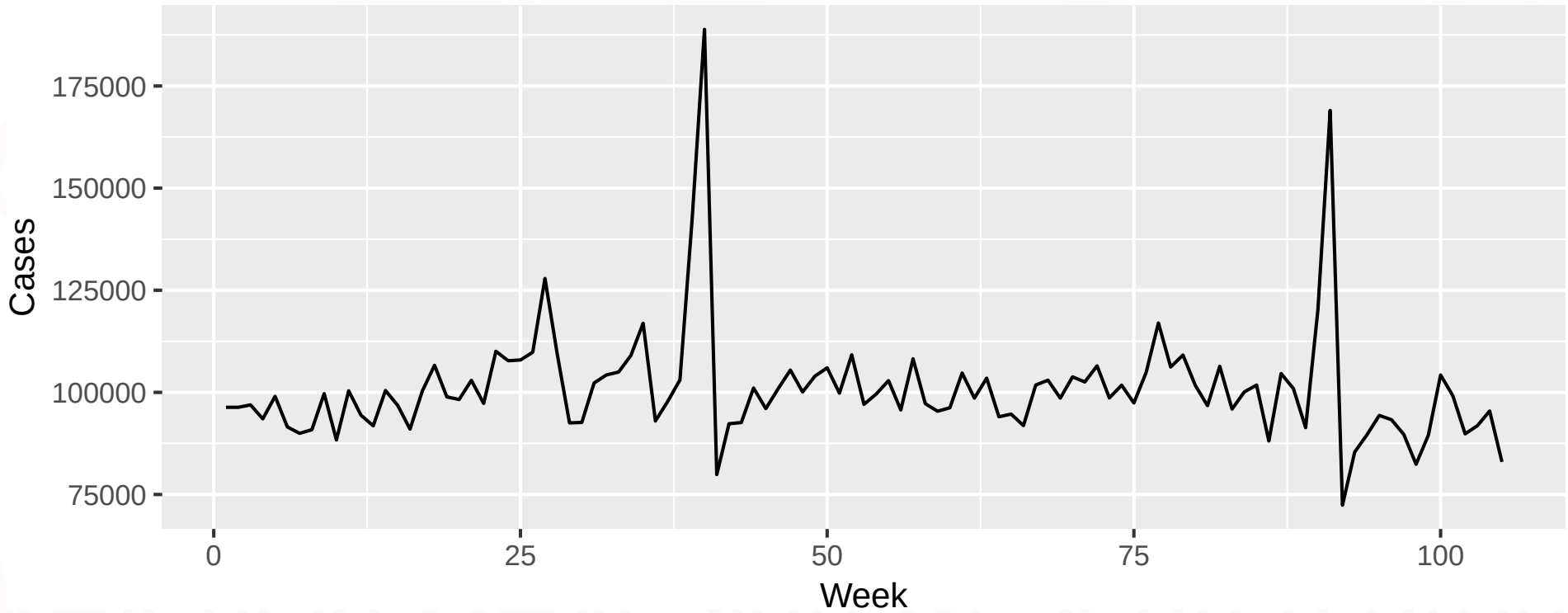


```
ggplot(Eggs, aes(y = Cases, x = Easter)) + geom_boxplot() + coord_fl
    labs(title = "BoxPlots por Pascua", y = "Cajas vendidas",
         x = "Mes")
```



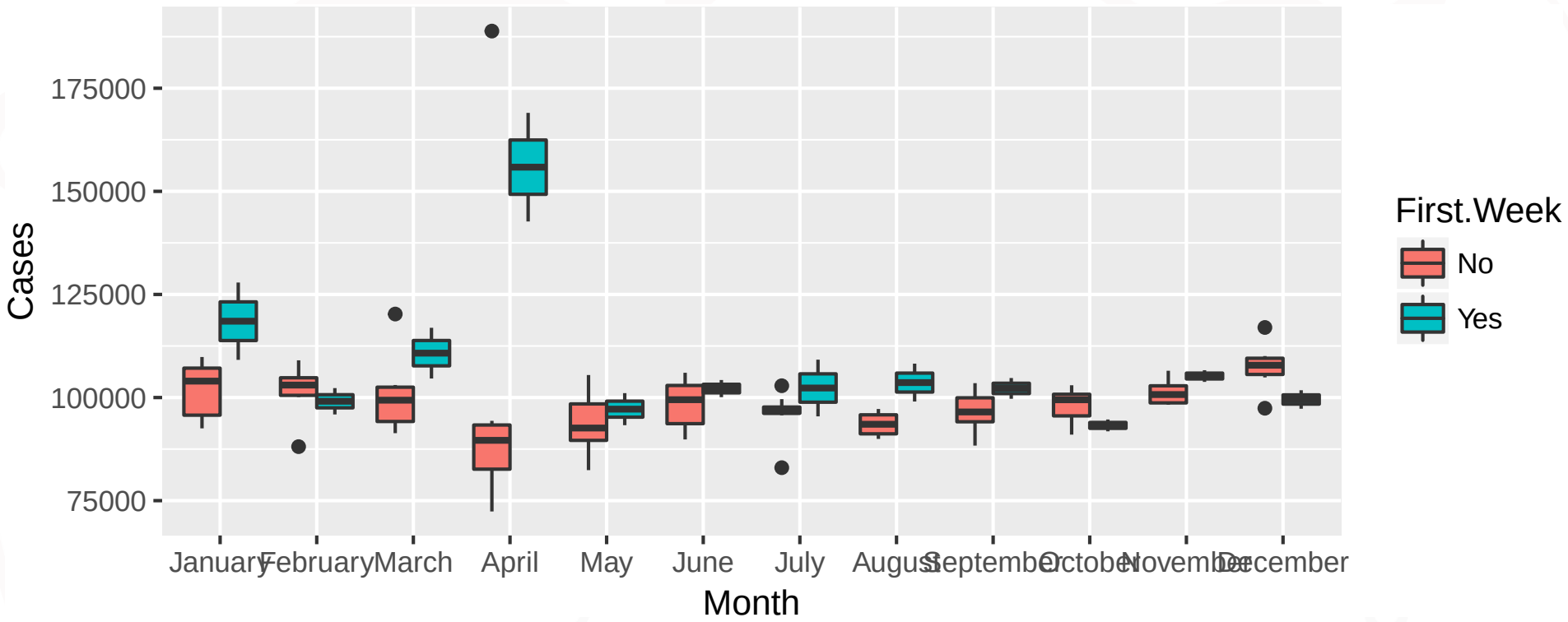
¿Qué pasa en aquí?

```
ggplot(Eggs, aes(y = Cases, x = Week)) + geom_line()
```

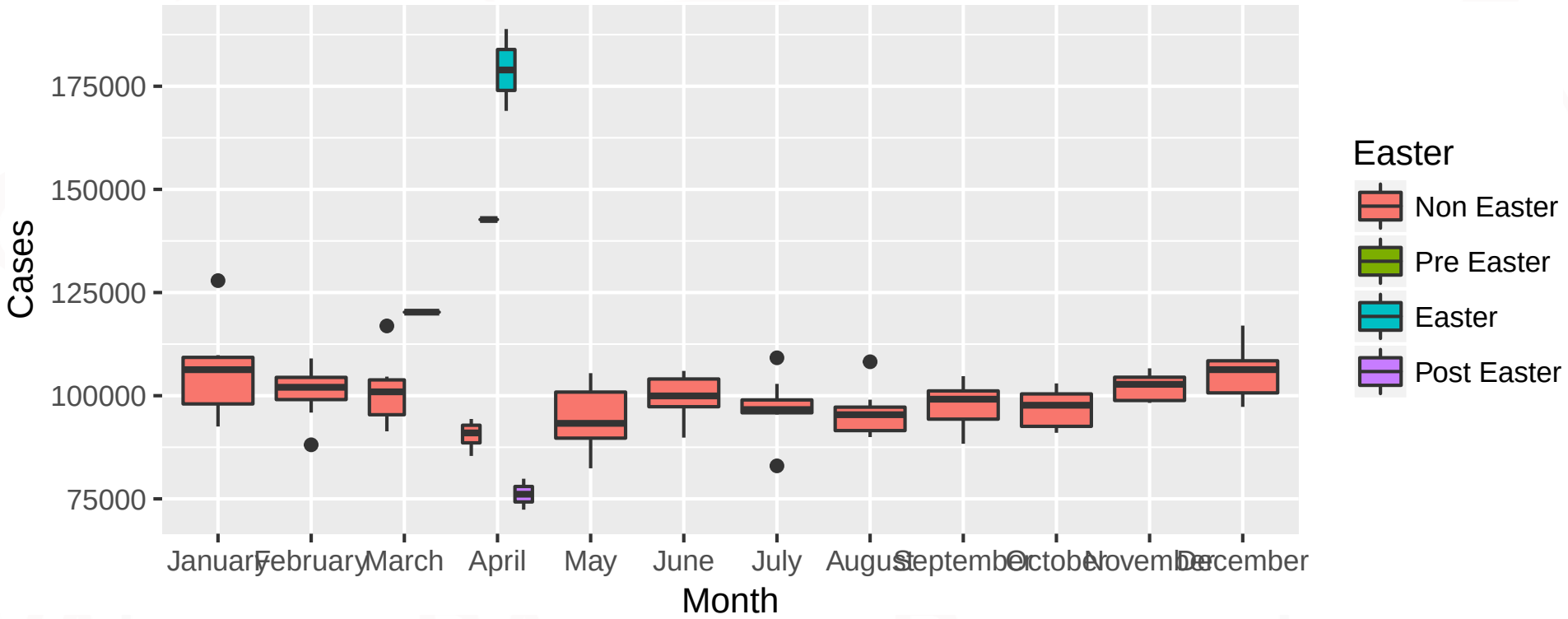


Agrupando por dos o más factores

```
ggplot(aes(y = Cases, x = Month, fill = First.Week), data = Eggs) +
    geom_boxplot()
```



```
ggplot(aes(y = Cases, x = Month, fill = Easter), data = Eggs) +
    geom_boxplot()
```



Planteando el modelo

Ya hemos visto el comportamiento de las variables "independientes" frente a la variable de estudio "**Cases**".

Ahora debemos pensar en cúal sería nuestro modelo de partida.

Vamos a diferencias 3 modelos:

- Uno solo con variables respuesta continuas ==> Regresión Lineal
- Uno solo con variables respuestas categoricas ==> ANOVA
- Uno que incluya las dos modelizacione santeriores ==> ANCOVA

Regresión Lineal Simple

Relaciones entre variables

Probando relaciones...

- ¿Modelo Más Complejo o Modelo más simple?
- ¿Introducir una a una las variables?
- ¿quitar una a una las variables has...?
- ¿Selección automática de Variables?

Relaciones entre variables

- Cuidado con de grados de libertad en la estimación de los parámetros y en el estudio de interacciones.
- Cuidado con la variable "Week", nos puede servir como una variable para estudiar la Temporalidad o estacionalidad de la "serie".
- La variable "Month" puede mostrar también la estacionalidad de la "serie".

Modelo completo (solo continuas)

```
fit1<- glm(Cases~Egg.Pr + Beef.Pr+
             Pork.Pr + Chicken.Pr + Cereal.Pr,
           data=Eggs, family="gaussian")

summary(fit1)
```

```
##
## Call:
## glm(formula = Cases ~ Egg.Pr + Beef.Pr + Pork.Pr + Chicken.Pr +
##     Cereal.Pr, family = "gaussian", data = Eggs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -28134   -6905   -1998    3914   73281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 231872.74   52571.74   4.411 2.62e-05 ***
## Egg.Pr        -624.83     197.83  -3.158   0.0021 **
## Beef.Pr        193.59     200.83   0.964   0.3374
## Pork.Pr        -58.45      95.32  -0.613   0.5412
## Chicken.Pr    -519.29     334.84  -1.551   0.1241
## Cereal.Pr     -487.28     311.52  -1.564   0.1210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 186806488)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 1.8494e+10  on  99  degrees of freedom
## AIC: 2305.6
##
## Number of Fisher Scoring iterations: 2
```

```
stargazer(fit1, header = FALSE, single.row = TRUE)
```

Table 2

	Dependent variable:
	Cases
Egg.Pr	−624.832*** (197.826)
Beef.Pr	193.593 (200.828)
Pork.Pr	−58.448 (95.324)
Chicken.Pr	−519.287 (334.836)
Cereal.Pr	−487.281 (311.520)
Constant	231,872.700*** (52,571.740)
Observations	105
Log Likelihood	−1,146.793
Akaike Inf. Crit.	2,305.585
Note:	*p<0.1; **p<0.05; ***p<0.01

Selección de modelos

- Esta es una línea de investigación que a día de hoy sigue abierta en Estadística.
- Existen diferentes criterios para comparar modelos: *AIC*, *BIC*, contrastes, etc.
- Y existen diferentes métodos de selección automática de variables: *Backward*, *Forward*, *Stepwise*.

Selcción de Modelos

- **Backward**: Introducir el modelo más complejo e ir eliminado una a una. La que sale ya no puede volver a entrar.
- **Forward**: Introducir una a una las variables hasta que ya no queden variables "significativas por meter""

Estos dos métodos tiene sus ventajas e invonvenientes, y no tienen porque coincidir el modelo que seleccionan cada una de ellas.

Selección de Modelos

- Ajustamos le modelo Completo

```
lmCompleto<-glm((Cases)~., data=Eggs[,5:10],
                family="gaussian")
summary(lmCompleto)
```

```
##
## Call:
## glm(formula = (Cases) ~ ., family = "gaussian", data = Eggs[,
##     5:10])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -28134   -6905   -1998    3914   73281
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 231872.74   52571.74   4.411 2.62e-05 ***
## Egg.Pr        -624.83     197.83  -3.158   0.0021 **
## Beef.Pr        193.59     200.83   0.964   0.3374
## Pork.Pr        -58.45      95.32  -0.613   0.5412
## Chicken.Pr    -519.29     334.84  -1.551   0.1241
## Cereal.Pr     -487.28     311.52  -1.564   0.1210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 186806488)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 1.8494e+10  on  99  degrees of freedom
## AIC: 2305.6
##
## Number of Fisher Scoring iterations: 2
```

- Ajustamos el modelo Nulo (solo β_0)

```
lmNulo <- glm(Cases ~ 1, data = Eggs[, 5:10], family = "gaussian")
summary(lmNulo)
```

```
##
## Call:
## glm(formula = Cases ~ 1, family = "gaussian", data = Eggs[, 5:10])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -28553   -6906   -1779    3328   87929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   100932       1398    72.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 205176880)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 2.1338e+10  on 104  degrees of freedom
## AIC: 2310.6
##
## Number of Fisher Scoring iterations: 2
```

Selección de Modelos

```
lmbac<-step(lmCompleto, method="backward")
lmfor<-step(lmNulo, scope=list(lower=lmNulo,
  upper=lmCompleto), direction="forward")
```

El modelo final que dice la selección es:

```
lmbac$formula
lmbac$coefficients
```

```
## Cases ~ Egg.Pr

## (Intercept)      Egg.Pr
## 153414.4947   -553.8708
```

```
lmStep <- step(lmCompleto, method = "both")
```

```
## (Cases) ~ Egg.Pr
```

```
## (Intercept)      Egg.Pr
## 153414.4947   -553.8708
```

Log(Cases)

- Ajustamos le modelo Completo

```
lmCompleto2<-glm(log(Cases)~.,
                 data=Eggs[,5:10], family="gaussian")
summary(lmCompleto2)
```

```
##
## Call:
## glm(formula = log(Cases) ~ ., family = "gaussian", data = Eggs[,
##     -1])
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.106712  -0.036544   0.002812   0.033391   0.121635
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.0919456  0.2544916  47.514  < 2e-16 ***
## MonthFebruary     -0.0217703  0.0267508  -0.814 0.418050
## MonthMarch        -0.0303768  0.0279841  -1.085 0.280806
## MonthApril        -0.1117185  0.0324185  -3.446 0.000890 ***
## MonthMay          -0.1257672  0.0276194  -4.554 1.77e-05 ***
## MonthJune         -0.0944817  0.0276486  -3.417 0.000977 ***
## MonthJuly         -0.1142706  0.0249947  -4.572 1.65e-05 ***
## MonthAugust       -0.1529920  0.0265341  -5.766 1.31e-07 ***
## MonthSeptember    -0.0831580  0.0273198  -3.044 0.003118 **
## MonthOctober      -0.0893844  0.0269269  -3.320 0.001335 **
## MonthNovember     -0.0294404  0.0263659  -1.117 0.267346
## MonthDecember      0.0169907  0.0272316   0.624 0.534362
## First.WeekYes      0.0557860  0.0132597   4.207 6.44e-05 ***
## EasterPre Easter   0.2602484  0.0439807   5.917 6.88e-08 ***
## EasterEaster       0.5345938  0.0500920  10.672  < 2e-16 ***
## EasterPost Easter -0.1932836  0.0472055  -4.095 9.69e-05 ***
## Egg.Pr            -0.0045553  0.0010801  -4.218 6.20e-05 ***
## Beef.Pr            0.0027199  0.0009152   2.972 0.003860 **
## Pork.Pr           -0.0002382  0.0004401  -0.541 0.589725
## Chicken.Pr        -0.0009578  0.0015190  -0.631 0.530030
```

```
lmStep2<-step(lmCompleto2, method="both")
```

```
## log(Cases) ~ Month + First.Week + Easter + Egg.Pr + Beef.Pr +
##     Cereal.Pr
```

```
##       (Intercept)     MonthFebruary        MonthMarch        MonthApril
##      11.997588382       -0.023079867       -0.031616978       -0.110683046
##          MonthMay         MonthJune          MonthJuly        MonthAugust
##      -0.120718086       -0.090220504       -0.114157113       -0.153155416
##    MonthSeptember      MonthOctober     MonthNovember     MonthDecember
##      -0.083942855       -0.087974072       -0.028132754        0.018149159
##     First.WeekYes  EasterPre Easter       EasterEaster EasterPost Easter
##       0.055153932        0.268538744        0.538095401       -0.195375762
##            Egg.Pr           Beef.Pr          Cereal.Pr
##      -0.004206962        0.002473655       -0.003464065
```

```
fit.5<-glm(log(Cases) ~ Month + First.Week +
             Easter + Egg.Pr + Beef.Pr +
             Cereal.Pr, data=Eggs, family="gaussian")

fit.6<-glm(log(Cases) ~ Egg.Pr +
             Cereal.Pr, data=Eggs,
                family="gaussian")

fit.6.1<-glm(log(Cases) ~ Egg.Pr +
               Cereal.Pr+ Easter,
                 data=Eggs, family="gaussian")
summary(fit.6.1)

library(lmtest)
bptest(fit.6)
shapiro.test(fit.6.1$residuals)
```

```
##
## Call:
## glm(formula = log(Cases) ~ Month + First.Week + Easter + Egg.Pr +
##     Beef.Pr + Cereal.Pr, family = "gaussian", data = Eggs)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.106213  -0.034748   0.003617   0.034333   0.126793
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.9975884  0.2183245  54.953  < 2e-16 ***
## MonthFebruary     -0.0230799  0.0259913  -0.888 0.377026
## MonthMarch        -0.0316170  0.0272474  -1.160 0.249111
## MonthApril        -0.1106830  0.0321301  -3.445 0.000886 ***
## MonthMay          -0.1207181  0.0266440  -4.531 1.88e-05 ***
## MonthJune         -0.0902205  0.0268349  -3.362 0.001156 **
## MonthJuly         -0.1141571  0.0246695  -4.627 1.30e-05 ***
## MonthAugust       -0.1531554  0.0261698  -5.852 8.59e-08 ***
## MonthSeptember    -0.0839429  0.0258198  -3.251 0.001642 **
## MonthOctober      -0.0879741  0.0246335  -3.571 0.000585 ***
## MonthNovember     -0.0281328  0.0256695  -1.096 0.276155
## MonthDecember      0.0181492  0.0266690   0.681 0.497995
## First.WeekYes      0.0551539  0.0131285   4.201 6.46e-05 ***
## EasterPre Easter   0.2685387  0.0420788   6.382 8.51e-09 ***
## EasterEaster       0.5380954  0.0494880  10.873  < 2e-16 ***
## EasterPost Easter -0.1953758  0.0463688  -4.214 6.18e-05 ***
## Egg.Pr            -0.0042070  0.0008716  -4.827 5.96e-06 ***
## Beef.Pr            0.0024737  0.0008146   3.037 0.003166 **
## Cereal.Pr         -0.0034641  0.0012859  -2.694 0.008490 **
## ---
```

```
##
## Call:
## glm(formula = log(Cases) ~ Egg.Pr + Cereal.Pr, family = "gaussian",
##     data = Eggs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.31614  -0.07054  -0.01571   0.04297   0.53632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.315449   0.287436  42.846  < 2e-16 ***
## Egg.Pr      -0.004589   0.001420  -3.231  0.00166 **
## Cereal.Pr   -0.003138   0.002225  -1.411  0.16140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.01334237)
##
##     Null deviance: 1.5325  on 104  degrees of freedom
## Residual deviance: 1.3609  on 102  degrees of freedom
## AIC: -150.33
##
## Number of Fisher Scoring iterations: 2
```

```
##
## Call:
## glm(formula = log(Cases) ~ Egg.Pr + Cereal.Pr + Easter, family = "gaussian",
##     data = Eggs)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.155604  -0.049442  -0.000084   0.051835   0.243255
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.0038926  0.1831178  65.553  < 2e-16 ***
## Egg.Pr            -0.0017032  0.0009272  -1.837   0.0692 .
## Cereal.Pr         -0.0029090  0.0013909  -2.091   0.0390 *
## EasterPre Easter   0.2935973  0.0515646   5.694 1.27e-07 ***
## EasterEaster       0.5559429  0.0535830  10.375  < 2e-16 ***
## EasterPost Easter -0.2528790  0.0516430  -4.897 3.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.00516773)
##
##     Null deviance: 1.53246  on 104  degrees of freedom
## Residual deviance: 0.51161  on  99  degrees of freedom
## AIC: -247.06
##
## Number of Fisher Scoring iterations: 2
```

```
library(lmtest)
bptest(fit.6)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit.6
## BP = 5.2689, df = 2, p-value = 0.07176
```

```
shapiro.test(fit.6.1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.6.1$residuals
## W = 0.98343, p-value = 0.2154
```

¿Es este el mejor modelo?

```
fit.step<-glm(Cases ~ Egg.Pr,
              family = "gaussian", data = Eggs[,5:10])
summary(fit.step)
```

```
##
## Call:
## glm(formula = Cases ~ Egg.Pr, family = "gaussian", data = Eggs[,
##     5:10])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -25831   -7104   -2765    3723   78333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 153414.5    15992.9   9.593 5.96e-16 ***
## Egg.Pr        -553.9      168.2  -3.293  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 187434583)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 1.9306e+10  on 103  degrees of freedom
## AIC: 2302.1
##
## Number of Fisher Scoring iterations: 2
```

ANOVA

Reajustando el Modelo

Ahora vamos a utilizar un mdoelo donde **TODAS** las variables independientes con FACTORES.

Es decir, Cases $\sim$ Month $+$ First.Week $+$ Easter.

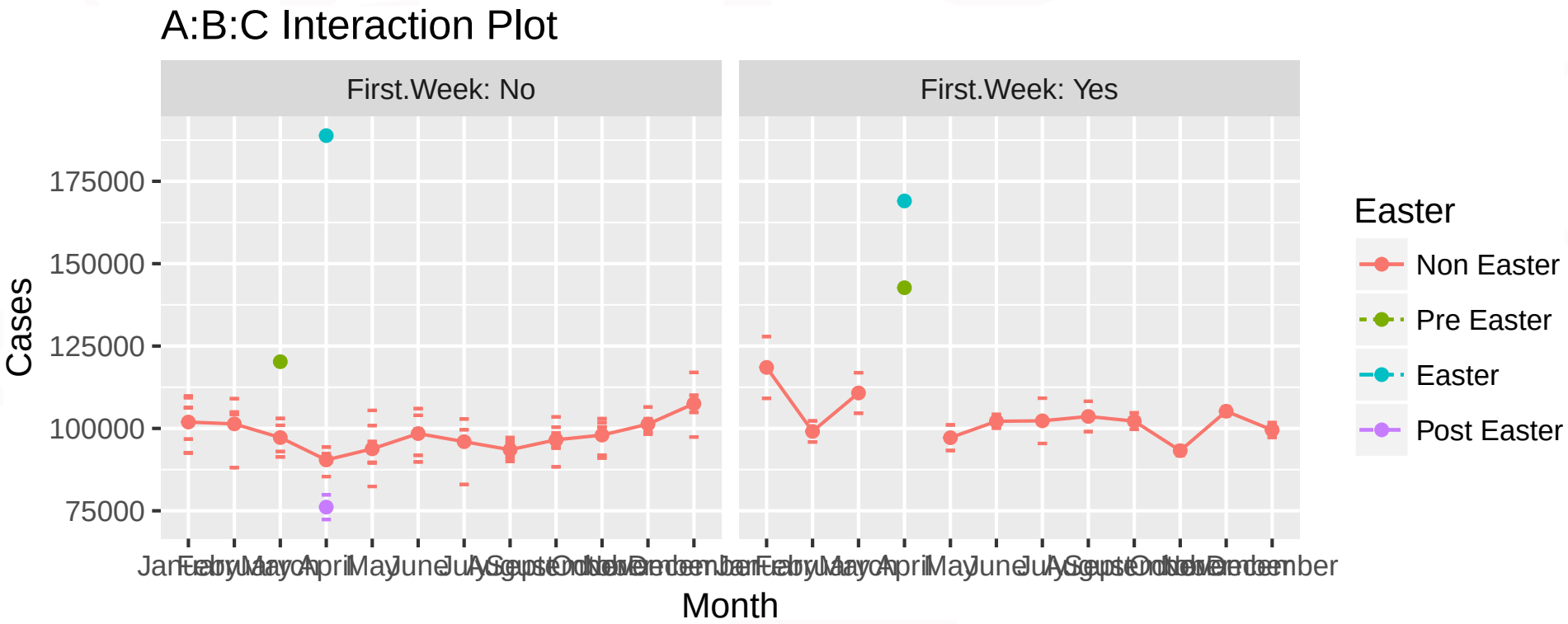
Es muy habitual que en estos modelos se consideren las **interacciones**, o sea, que el comportamiento de lso factores entre si no es homogéneo.

Por tanto el modelo quedaría como:

Cases $\sim$ Month $\times$ First.Week $\times$ Easter

Gráfico de interacciones triples

```
library(dae)
interaction.ABC.plot(Cases,  Month ,Easter,First.Week,
                     data=Eggs,
                     ggplotFunc=
                       list(geom_errorbar(
                           data=Eggs,
                           aes(ymax=Cases, ymin=Cases),
                                                width=0.2)))
```

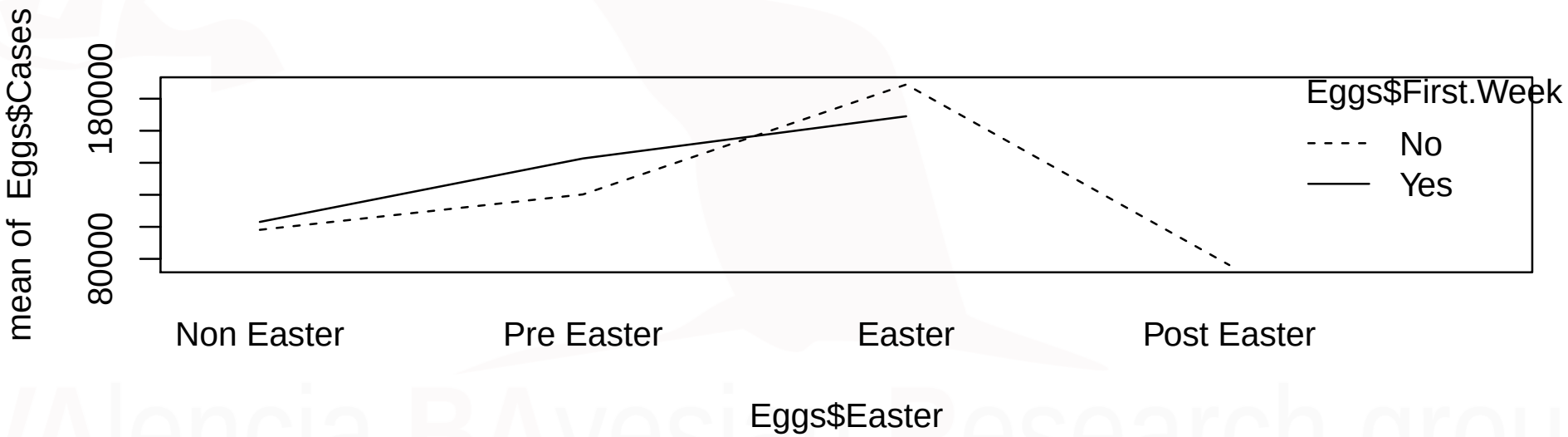


Interacción de dos Factores

Una forma de hacerlo muy bonito: Enlace

Y una form amáinteraction.plot(dose, supp, len, fixed = TRUE) s básica:

```
interaction.plot(Eggs$Easter, Eggs$First.Week, Eggs$Cases, fixed = T
```



Ajustando el modelo

```
##
## Call:
## glm(formula = Cases ~ Month + First.Week * Easter, family = "gaussian",
##     data = Eggs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13377.7   -3745.0     404.8    3474.3   19123.7
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  104368.05    2030.17  51.409  < 2e-16 ***
## MonthFebruary                 -4660.86    3010.71  -1.548 0.125229
## MonthMarch                    -4533.98    3129.68  -1.449 0.151015
## MonthApril                   -13941.05    3766.49  -3.701 0.000375 ***
## MonthMay                     -10794.11    2915.55  -3.702 0.000374 ***
## MonthJune                     -6096.74    3010.71  -2.025 0.045931 *
## MonthJuly                     -8002.40    2837.59  -2.820 0.005944 **
## MonthAugust                   -9569.99    2915.55  -3.282 0.001483 **
## MonthSeptember                -7506.61    3010.71  -2.493 0.014549 *
## MonthOctober                  -8216.50    2837.59  -2.896 0.004784 **
## MonthNovember                 -3190.99    3010.71  -1.060 0.292133
## MonthDecember                    38.89    3010.71   0.013 0.989724
## First.WeekYes                  4395.25    1546.21   2.843 0.005575 **
## EasterPre Easter              20412.93    6797.50   3.003 0.003490 **
## EasterEaster                  98434.00    7093.97  13.876  < 2e-16 ***
## EasterPost Easter            -14303.00    5494.96  -2.603 0.010865 *
## First.WeekYes:EasterPre Easter  27458.82   10014.35   2.742 0.007413 **
## First.WeekYes:EasterEaster    -24228.25    9105.48  -2.661 0.009281 **
## First.WeekYes:EasterPost Easter       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 40259482)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 3.5026e+09  on  87  degrees of freedom
```

re-Ajsutando el modelo

```
##
## Call:
## glm(formula = Cases ~ Month + First.Week + Easter, family = "gaussian",
##     data = Eggs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13348.4   -3906.5     591.6    3558.2   19006.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       104338.77    2172.86  48.019  < 2e-16 ***
## MonthFebruary      -4668.18    3224.73  -1.448 0.151236
## MonthMarch         -6174.62    3289.91  -1.877 0.063817 .
## MonthApril        -11062.60    3869.19  -2.859 0.005291 **
## MonthMay          -10797.36    3122.85  -3.458 0.000838 ***
## MonthJune          -6104.06    3224.73  -1.893 0.061622 .
## MonthJuly          -8002.40    3039.36  -2.633 0.009981 **
## MonthAugust        -9573.25    3122.85  -3.066 0.002877 **
## MonthSeptember     -7513.93    3224.73  -2.330 0.022063 *
## MonthOctober       -8216.50    3039.36  -2.703 0.008222 **
## MonthNovember      -3198.31    3224.73  -0.992 0.323980
## MonthDecember         31.57    3224.73   0.010 0.992212
## First.WeekYes       4541.66    1600.53   2.838 0.005629 **
## EasterPre Easter   33479.51    5277.45   6.344 9.11e-09 ***
## EasterEaster       83397.50    5811.85  14.350  < 2e-16 ***
## EasterPost Easter -17152.17    5786.30  -2.964 0.003893 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 46188668)
```

```
##
## Call:
## glm(formula = Cases ~ First.Week + Easter, family = "gaussian",
##     data = Eggs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -15789.8   -5582.8    -408.8    4755.2   24964.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        98188.8      830.8 118.179  < 2e-16 ***
## First.WeekYes       4733.9     1723.7   2.746  0.00715 **
## EasterPre Easter   30914.8     5261.2   5.876 5.55e-08 ***
## EasterEaster       78388.8     5261.2  14.900  < 2e-16 ***
## EasterPost Easter -22064.8     5253.3  -4.200 5.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 53813783)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 5.3814e+09  on 100  degrees of freedom
## AIC: 2174
##
## Number of Fisher Scoring iterations: 2
```

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ First.Week + Easter
## Model 2: Cases ~ Month + First.Week + Easter
##   Resid. Df Resid. Dev Df   Deviance
## 1       100 5381378312
## 2        89 4110791433 11 1270586879
```

```
##
## Call:
## glm(formula = Cases ~ First.Week + Easter + First.Week:Easter,
##     family = "gaussian", data = Eggs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -15745   -5152       0    4411   24808
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     98144.3      807.6 121.521  < 2e-16 ***
## First.WeekYes                    4934.2     1713.3   2.880  0.00488 **
## EasterPre Easter                22102.7     7132.8   3.099  0.00254 **
## EasterEaster                    90716.7     7132.8  12.718  < 2e-16 ***
## EasterPost Easter              -22020.3     5075.9  -4.338 3.49e-05 ***
## First.WeekYes:EasterPre Easter  17512.8    10167.9   1.722  0.08816 .
## First.WeekYes:EasterEaster     -24767.2    10167.9  -2.436  0.01666 *
## First.WeekYes:EasterPost Easter       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 50224986)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 4.9220e+09  on  98  degrees of freedom
## AIC: 2168.6
##
## Number of Fisher Scoring iterations: 2

##
```

Niveles no significativos, ¿qué hacer?

¡NUNCA debemos eliminar niveles de una variable FACTOR!

- Podemos reagrupar (`ifelse`)
- Podemos comprobar si una variable aporta o no en globar al modelo (reducción de la Deviance)

```
anova(fit.anova2.1, fit.anova2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ First.Week + Easter
## Model 2: Cases ~ First.Week + Easter + First.Week:Easter
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1       100 5381378312
## 2        98 4922048604  2 459329707  0.01033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit.anova3 <- glm(Cases ~ Month + Easter, data = Eggs, family = "gaussian")
anova(fit.anova3, fit.anova2, test = "LRT")
```

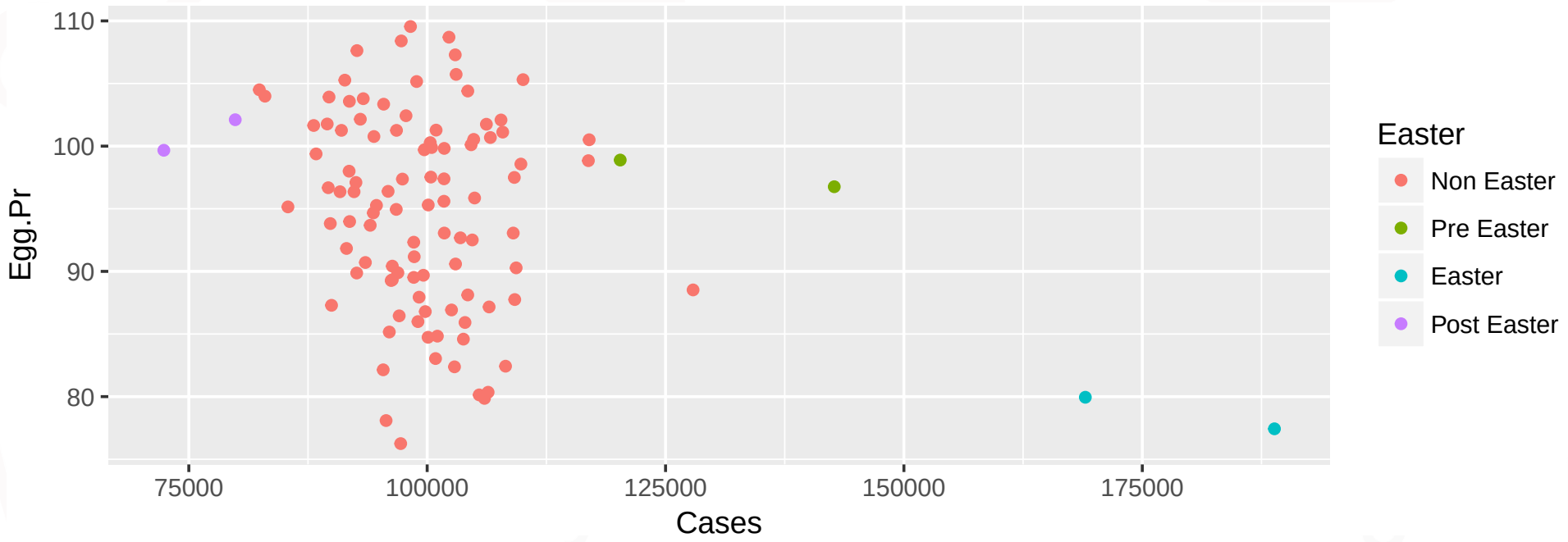
```
## Analysis of Deviance Table
##
## Model 1: Cases ~ Month + Easter
## Model 2: Cases ~ First.Week + Easter + First.Week:Easter
##   Resid. Df Resid. Dev Df   Deviance Pr(>Chi)
## 1        90 4482701026
## 2        98 4922048604 -8 -439347578   0.3576
```

ANCOVA

Variables contínuas y Factores como independientes

Este es el caso de la vida real, y además, suelen existir relaciones entre lso factores y las contínuas.

```
ggplot(Eggs, aes(Cases, Egg.Pr, color = Easter)) + geom_point()
```



El mdoelo Completo

- No simepre es posible realizar un ajuste de un modelo completo.
- En este caso sería un modelo donde cada X-contínua está multiplicada por el factor:
  - Egg.Pr*Easter* + *Egg.Pr*First.Week + Egg.Pr*Month

Probemos este modelo: Cases $\sim$ Egg.Pr*Easter* + *Egg.Pr*First.Week + Egg.Pr*Month + Beef.Pr + . . . + Cereal.Pr

```
fit.completo1<-glm(Cases~ Egg.Pr*Easter +
                     Egg.Pr*First.Week +
                     Egg.Pr*Month  + Beef.Pr +
                     Pork.Pr+  Chicken.Pr +
                     Cereal.Pr,
                   family='gaussian', data=Eggs)
summary(fit.completo1)
```

```
##
## Call:
## glm(formula = Cases ~ Egg.Pr * Easter + Egg.Pr * First.Week +
##     Egg.Pr * Month + Beef.Pr + Pork.Pr + Chicken.Pr + Cereal.Pr,
##     family = "gaussian", data = Eggs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -10691   -2376       0    3050   11753
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                178003.35   33218.64   5.359 1.05e-06 ***
## Egg.Pr                       -708.70     229.20  -3.092 0.002867 **
## EasterPre Easter          1639284.93  442507.82   3.705 0.000423 ***
## EasterEaster               688044.17  387770.31   1.774 0.080415 .
## EasterPost Easter         -604850.68  442098.41  -1.368 0.175708
## First.WeekYes              35785.65   16600.19   2.156 0.034594 *
## MonthFebruary             -51460.61   44493.69  -1.157 0.251432
## MonthMarch                155053.33   90775.85   1.708 0.092115 .
## MonthApril                 57766.39  297873.29   0.194 0.846801
## MonthMay                  -46423.52   29140.97  -1.593 0.115716
## MonthJune                 -61218.05   37366.50  -1.638 0.105909
## MonthJuly                 -58410.40   31787.13  -1.838 0.070435 .
## MonthAugust               -51024.55   36440.48  -1.400 0.165929
## MonthSeptember             59921.73   61367.76   0.976 0.332259
## MonthOctober                -289.63   45076.82  -0.006 0.994892
## MonthNovember             -57577.59   29204.39  -1.972 0.052673 .
## MonthDecember             -47817.66   60523.78  -0.790 0.432197
## Beef.Pr                      355.09      94.99   3.738 0.000379 ***
## Pork.Pr                      -25.54      48.62  -0.525 0.601009
## Chicken.Pr                  -104.27     161.23  -0.647 0.519990
## Cereal.Pr                   -399.17     190.36  -2.097 0.039668 *
## Egg.Pr:EasterPre Easter   -16473.33    4519.11  -3.645 0.000514 ***
## Egg.Pr:EasterEaster        -8023.07    4428.39  -1.812 0.074379 .
## Egg.Pr:EasterPost Easter    5902.57    4489.73   1.315 0.192969
## Egg.Pr:First.WeekYes        -324.73     175.22  -1.853 0.068125 .
## Egg.Pr:MonthFebruary         510.07     452.23   1.128 0.263270
## Egg.Pr:MonthMarch          -1515.48     891.15  -1.701 0.093524 .
## Egg.Pr:MonthApril           -745.52    3111.10  -0.240 0.811326
## Egg.Pr:MonthMay              352.00     310.34   1.134 0.260625
## Egg.Pr:MonthJune             549.43     410.39   1.339 0.185034
```

Selecciones automáticamente

```
library(MASS)
fit.completo.step <- stepAIC(fit.completo1, direcction = "both")
```

El 'mejor' modelo es:

```
## Cases ~ Egg.Pr + Easter + First.Week + Month + Beef.Pr + Cereal.P
##     Egg.Pr:Easter
```

re-Ajustando

```
fit.completo2<-glm(Cases ~ Egg.Pr +
                     Easter + First.Week +
                     Month + Beef.Pr +
                     Cereal.Pr +
                     Egg.Pr:Easter,
                     data=Eggs, family="gaussian")
summary(fit.completo2)
```

```
##
## Call:
## glm(formula = Cases ~ Egg.Pr + Easter + First.Week + Month +
##     Beef.Pr + Cereal.Pr + Egg.Pr:Easter, family = "gaussian",
##     data = Eggs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10959.0   -3197.5       0.4    2940.8   15620.0
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                141694.03   21346.55   6.638 3.07e-09 ***
## Egg.Pr                       -420.27      83.58  -5.028 2.81e-06 ***
## EasterPre Easter          1288643.10  379463.63   3.396 0.001052 **
## EasterEaster               714789.67  232056.32   3.080 0.002805 **
## EasterPost Easter         -476819.01  304606.42  -1.565 0.121303
## First.WeekYes                5296.56    1294.16   4.093 9.84e-05 ***
## MonthFebruary               -2812.35    2490.22  -1.129 0.262001
## MonthMarch                  -2239.18    2648.38  -0.845 0.400267
## MonthApril                 -13785.33    3250.04  -4.242 5.74e-05 ***
## MonthMay                   -12556.06    2559.10  -4.906 4.56e-06 ***
## MonthJune                   -9632.83    2572.86  -3.744 0.000333 ***
## MonthJuly                  -11712.31    2362.50  -4.958 3.72e-06 ***
## MonthAugust                -15501.99    2507.50  -6.182 2.26e-08 ***
## MonthSeptember              -8617.15    2474.36  -3.483 0.000795 ***
## MonthOctober                -8959.11    2363.47  -3.791 0.000284 ***
## MonthNovember               -3136.47    2459.71  -1.275 0.205818
## MonthDecember                1519.74    2554.06   0.595 0.553444
## Beef.Pr                       250.63      78.08   3.210 0.001888 **
## Cereal.Pr                    -287.49     127.88  -2.248 0.027220 *
## Egg.Pr:EasterPre Easter   -12850.86    3876.13  -3.315 0.001359 **
## Egg.Pr:EasterEaster        -8104.30    2947.39  -2.750 0.007320 **
## Egg.Pr:EasterPost Easter    4587.81    3017.59   1.520 0.132222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 26355714)
##
##     Null deviance: 2.1338e+10  on 104  degrees of freedom
## Residual deviance: 2.1875e+09  on  83  degrees of freedom
## AIC: 2113.4
```

Interpretando los Betas

- Se trata de ver cómo varía la venta de cajas por aumentos o disminuciones de las variables.
- Recordar que los factores, el nivel de referencia o más bajo está siempre en el intercept incluido.

Table 3

	Dependent variable:
	Cases
Egg.Pr	−420.274*** (−584.095, −256.453)
EasterPre Easter	1,288,643.000*** (544,908.100, 2,032,378.000)
EasterEaster	714,789.700*** (259,967.600, 1,169,612.000)
EasterPost Easter	−476,819.000 (−1,073,837.000, 120,198.600)
First.WeekYes	5,296.564*** (2,760.051, 7,833.076)
MonthFebruary	−2,812.351 (−7,693.099, 2,068.398)
MonthMarch	−2,239.181 (−7,429.904, 2,951.542)
MonthApril	−13,785.330*** (−20,155.280, −7,415.376)
MonthMay	−12,556.060*** (−17,571.810, −7,540.308)
MonthJune	−9,632.826*** (−14,675.540, −4,590.117)
MonthJuly	−11,712.310*** (−16,342.740, −7,081.891)
MonthAugust	−15,501.990*** (−20,416.610, −10,587.380)
MonthSeptember	−8,617.148*** (−13,466.810, −3,767.482)
MonthOctober	−8,959.110*** (−13,591.440, −4,326.784)
MonthNovember	−3,136.469 (−7,957.408, 1,684.469)
MonthDecember	1,519.738 (−3,486.129, 6,525.605)
Beef.Pr	250.632*** (97.603, 403.662)
Cereal.Pr	−287.487** (−538.126, −36.848)
Egg.Pr:EasterPre Easter	−12,850.860*** (−20,447.930, −5,253.784)
Egg.Pr:EasterEaster	−8,104.300*** (−13,881.080, −2,327.524)
Egg.Pr:EasterPost Easter	4,587.811 (−1,326.555, 10,502.180)
Constant	141,694.000*** (99,855.570, 183,532.500)
Observations	105
Akaike Inf. Crit.	2,113.445
Note:	*p<0.1; **p<0.05; ***p<0.01