Trabalho 2

Daniel Krügel

2023-05-14

Questão 1

Para o ajuste do modelo coppula segue o código e o sumário da regressão:

```
fitcoppula <- gcmr(status ~ treat + age + type + obs_time,
     marginal = gaussian.marg(),
     cormat = cluster.cormat(id, type = "exchangeable"),
     data = dados)

summary(fitcoppula)
```

```
##
## Call:
## gcmr(formula = status ~ treat + age + type + obs_time, data = dados,
##     marginal = gaussian.marg(), cormat = cluster.cormat(id, type = "exchangeable"))
##
##
## Coefficients marginal model:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.9093338  0.0867490  10.482   <2e-16 ***
## treat       -0.0227774  0.0180360  -1.263    0.207
## age         -0.0013458  0.0027967  -0.481    0.630
## type         0.0876196  0.0838016   1.046    0.296
## obs_time    -0.0167374  0.0009691 -17.272   <2e-16 ***
## sigma        0.3774566  0.0148840  25.360   <2e-16 ***
##
## Coefficients Gaussian copula:
##     Estimate Std. Error z value Pr(>|z|)
## tau   0.4185     0.0632   6.622 3.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## log likelihood = 156.22,  AIC = 326.44
```

Aparentemente a unica variável importante para o modelo é o tempo desde o tratamento. Como a Marginal foi informada pelo trabalho, não testei com outras para não me extender porém escolhi o melhor ajuste utilizando o critério de Akaike (AIC) já que para os modelos Coppula os resíduos do modelo não são informativos sobre a correção da gaussiana de coppula.

Questão 2

Para a questão 2 temos que análisar os dados porém sem um modelo em mente, aproveitei para aplicar modelos que já foram vistos, incluindo modelos lineares simples, regressão logistica e modelos de regressão categóricas. Nenhuma dessas foi de longe informativa sobre a modelagem dos dados. Portanto após uma

análise descritiva dos dados a ficha me caiu e fiz uma regressão generalizada de poisson, já que os dados da variável resposta compreendem a uma resposta de contagem.

```
fitglm <- glm(y ~.,
              data = dados,
              family = poisson)
car::Anova(fitglm)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##        LR Chisq Df Pr(>Chisq)
## Price    0.8910  1    0.34520
## Gender   0.0049  1    0.94447
## Brand    5.1899  1    0.02272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A unica variável significativa neste modelo foi a Brand, representando a marca do cigarro.

```
fitglm2 <- glm(y ~. + Brand:Price,
               family = poisson,
               data = dados)
car::Anova(fitglm2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##             LR Chisq Df Pr(>Chisq)
## Price         0.8910  1    0.34520
## Gender        1.0735  1    0.30015
## Brand         5.1899  1    0.02272 *
## Price:Brand   3.9500  1    0.04687 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O modelo com interação entre o preço e a marca do cigarro aparenta ser significativa também.

```
fitglm3 <- glm(y ~. + I(Price^2), family = poisson, data = dados)
car::Anova(fitglm3)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##            LR Chisq Df Pr(>Chisq)
## Price       1.36166  1     0.2433
## Gender      0.71672  1     0.3972
## Brand       0.24768  1     0.6187
## I(Price^2)  1.54525  1     0.2138
```

O modelo com a adição do modelo quadrático para preço perdeu um pouco do poder de explicação das variáveis dependentes.

Para comparar os 3 podemos usar novamente o critério de akaike para desempate:

```
data.frame("Simples" = AIC(fitglm), "Interação" = AIC(fitglm2), "Quadratico"=AIC(fitglm3))
```

```
##    Simples Interação Quadratico
## 1 105.5631  103.6131   106.0178
```

O modelo de interação apresentou um menor AIC, portanto é o modelo mais parcimonioso entre os 3.

Para realizarmos a regressão Ridge precisamos criar as matrizes do modelo:

```
dados$Gender <- ifelse(dados$Gender == "M", 1,0)
dados$Brand <- ifelse(dados$Brand == "A",0,1)
y <- dados$y
x <- as.matrix(dados[,-1])
```

Em seguida podemos usar a função `glmnet` e utilizar a função `cv.glmnet` para calcular o Lambda ideal utilizando validação cruzada:

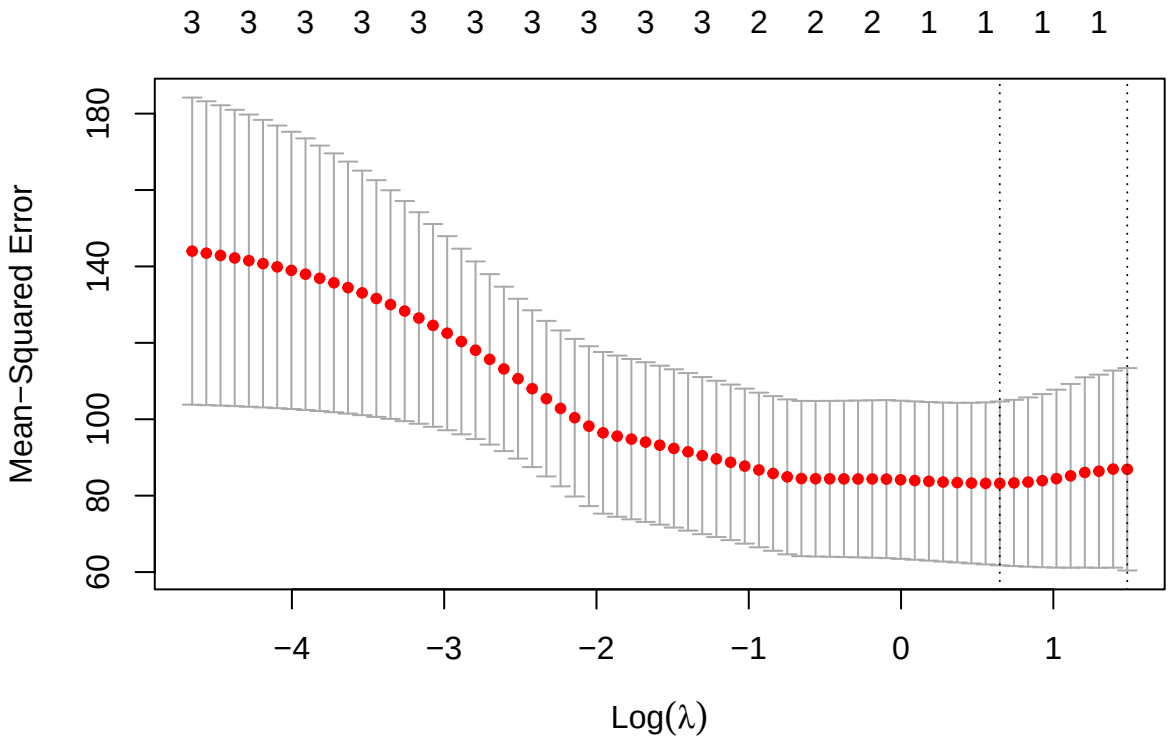
```
regRidge <- glmnet(x,y,
                   family = "poisson",
                   alpha = 0)

LambdaCalculado <- cv.glmnet(x,y)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```
regRidge <- glmnet(x,y,
                   family = "poisson",
                   alpha = 0,
                   lambda = LambdaCalculado$lambda.min)
```

```
plot(LambdaCalculado)
```



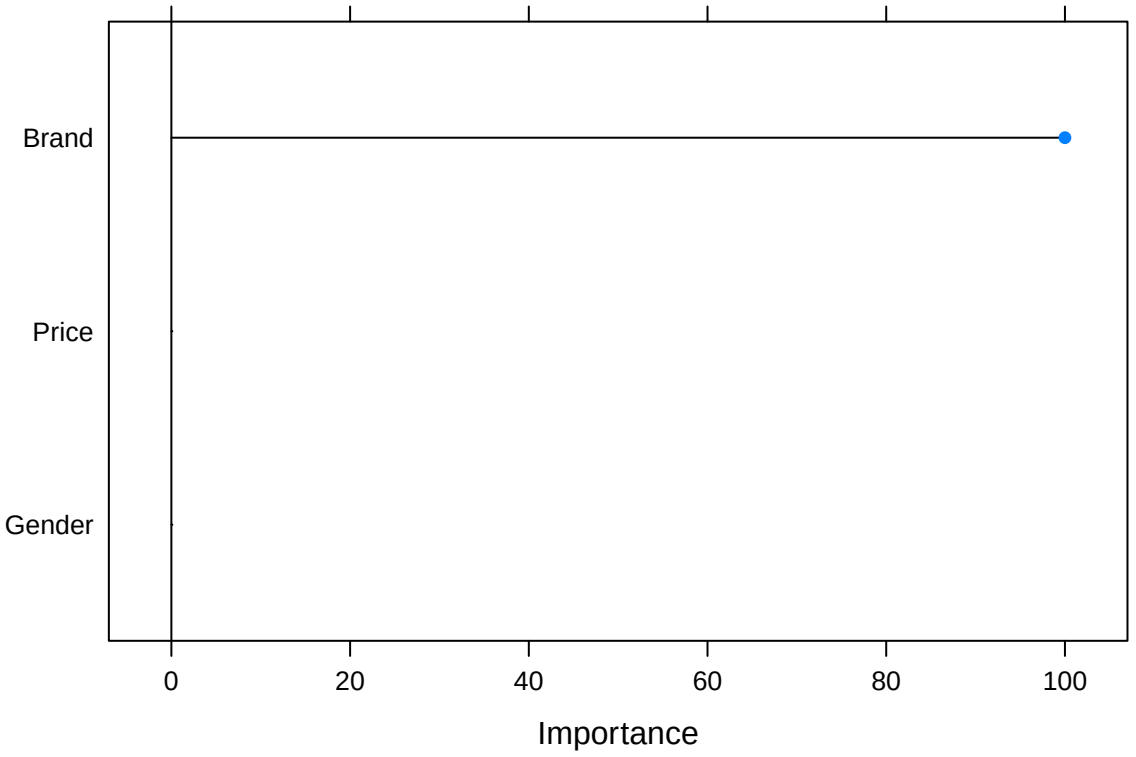
No ultimo gráfico podemos dar uma olhada nos diferentes valores de Lambda possíveis versus o erro quadrático médio, uma visualização que nos ajuda a entender a importância de escolher um lambda ótimo.

Para olharmos quais as variáveis de mais importância para a regressão vamos usar o pacote `caret` para treinar um modelo:

```r
library(caret)
set.seed(123)
modelo <- train(
  form = y ~.,
  method = "glmnet",
  data = dados,
  lambda = LambdaCalculado$lambda.min
)
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.
```

```r
plot(varImp(modelo))
```



Contrariando os outros metodos, utilizando Regressão ridge encontramos que a variável Price é a mais influente.