Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

EXPLORANDO OS MODELOS LINEARES GENERALIZADOS APLICAÇÃO A DADOS DE UM PEQUENO SUPERMERCADO

CE225 - Modelos Lineares Generalizados

Eduardo Elias Ribeiro Junior

Curitiba, 17 de novembro de 2014

Contents

1 Introdução

Para aplicação dos conceitos apresentados durante a disciplina de Modelos Lineares Generalizados foi disponibilizado um conjunto de dados com 100 observações referentes ao gasto de clientes de um pequeno supermercado. Neste conjunto de dados foram coletadas as informações de *forma de pagamento*, *tipo de cliente*, *Distância até o supermercado*, *Número de pessoas que moram com o cliente* e *Valor gasto na compra*, neste trabalho estas variáveis serão nomeadas como *X1*, *X2*, *X3*, *X4* e *Gasto* respectivamente. Abaixo temos detalhadas as variáveis:

- **X1**: Forma de pagamento da compra (Variável Categórica).
  $x1$ - Dinheiro, Cartão de Crédito ou Vale Alimentação;
- **X2**: Tipo de cliente (Variável Categórica).
  $x2$ - Cliente cadastrado ou Cliente Não Cadastrado;
- **X3**: Distância entre a residência do cliente e o supermercado (Variável numérica).
  $x3 \in R_+$ em km.
- **X4**: Número de pessoas que moram com o cliente, incluindo o próprio cliente (Variável numérica).
  $x4 \in Z_+^*$.
- **Gasto**: Gasto do cliente em sua última compra (Variável numérica).
  $Gasto \in R_+$ em centenas de reais.

Com esse conjunto de dados deseja explicar a variável *Gasto* com base nas demais variáveis a partir de um modelo linear generalizado, cujo especificações estarão descritas nas próximas seções.

2 Modelagem

Nesta seção apresentaremos e discutiremos os principais tópicos para modelagem de dados considerando um modelo linear generalizado.

Abaixo são exibidas as 10 primeiras observações contidas na base de dados, cujo total de observações é 100.

Table 1: Estrutura da base de dados

X1	X2	X3	X4	Gasto
Vale Alimentação	Cliente cadastrado	0.30	1	0.35
Vale Alimentação	Cliente não cadastrado	2.70	1	0.32
Cartão de crédito	Cliente cadastrado	9.90	3	1.12
Vale Alimentação	Cliente não cadastrado	3.20	1	0.69
Vale Alimentação	Cliente não cadastrado	10.00	4	0.46
Vale Alimentação	Cliente cadastrado	7.30	6	0.76
Cartão de crédito	Cliente não cadastrado	3.30	9	1.62
Dinheiro	Cliente não cadastrado	9.00	3	0.82
Dinheiro	Cliente não cadastrado	0.30	3	0.09
Cartão de crédito	Cliente não cadastrado	4.00	6	0.32

2.1 Análise Descritiva e Exploratória

Esta etapa, preliminar do processo de modelagem, é de extrema importância para que se especifique um bom modelo de acordo com as indicações a serem observadas.

Primeiramente estudaremos o comportamento da variável resposta.

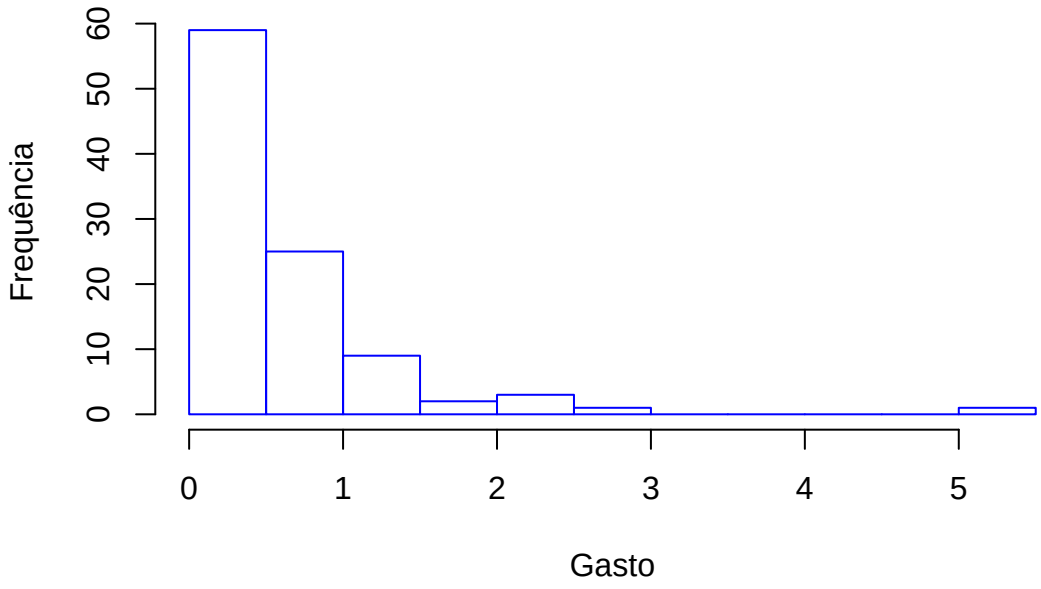

Figure 1: Histograma de Gasto

Table 2: Medidas Descritivas para Gasto

Medidas	Valores
Min.	0.06
1st Qu.	0.23
Mediana	0.40
Média	0.60
3rd Qu.	0.68
Max.	5.22
Variância	0.47
Desvio Padrão	0.69
Coeficiente de Variação	1.14

Perceba que tanto pela Figura 1 quanto pela Tabela 1 é evidente a assimetria a direita da distribuição da variável *Gasto*, ficando mais de 80% das observações entre 0 e 1 centenas de reais.

Agora estudando a variável resposta em função das variáveis explicativas, como temos variáveis categóricas e numéricas serão apresentados gráficos de caixas (box-plots) para as variáveis categóricas e gráficos de dispersão (scatter-plots) para as variáveis numéricas.

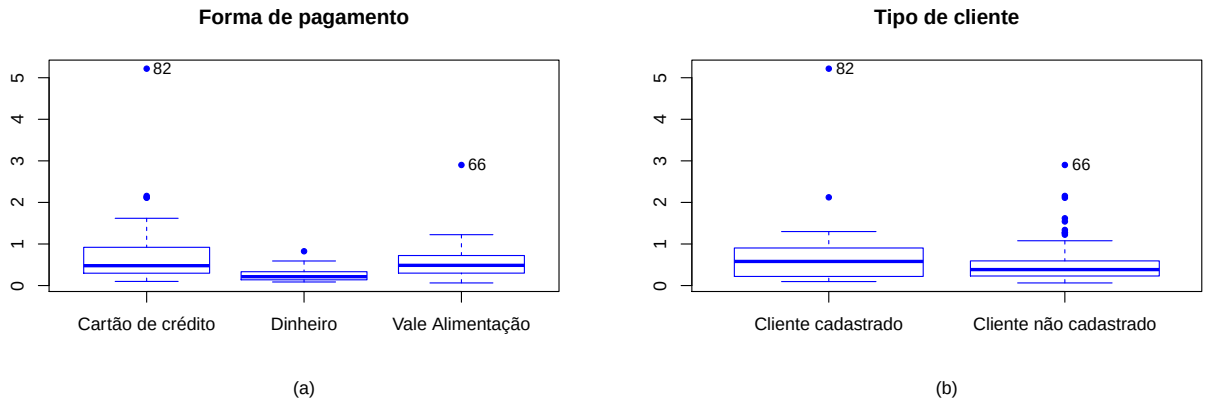

Figure 2: Box-plots de Gasto em relação a X1 e X2

Oberve na Figura 2 (a) que o gasto de clientes que utilizam cartão de crédito como forma de pagamento é maior do que que as que utilizam dinheiro e tem seu comportamento bem parecido com o gasto dos clientes que utilizam vale alimentação, apresentando um gasto levemente superior com variabilidade maior em relação a esta categoria. Ainda pode-se notar que os clientes que optam pagar com dinheiro tendem a ter um gasto menor e ainda com menor variabilidade, estando todos os indíviduos, nesta categoria, com gasto entre 0 e 1 centenas de reais. Já na Figura 2 (b) o compartamento da vairiável gasto entre as duas categorias, cliente não cadastrado e cliente cadastrado, é relativamente parecido e parece que se tem uma menor dispersão de valores gastos para clientes não cadastrados, porém perceba que há várias observações que extrapolam o limite superior do box-plot.

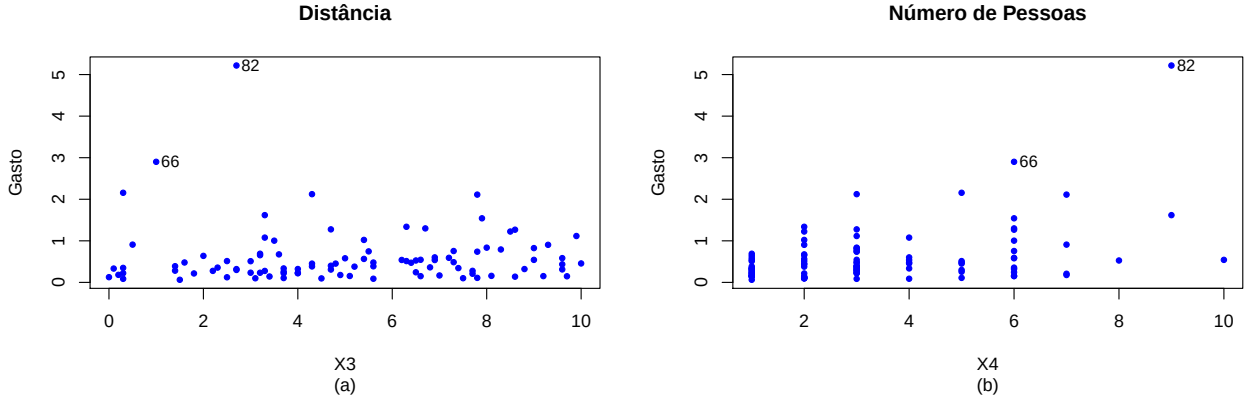

Figure 3: Scatter-plots de Gasto em relação a X3 e X4

Na Figura 3 observamos que em ambos os gráficos não temos uma forte relação (crescente ou decrescente) entre a variável resposta e as variáveis explicativas numéricas, porém perceba que a dispersão da variável resposta não parece ser constante entre os valores das variáveis explicativas, na Figura 3 (b) esta variação não constante é mais visível.

Como complemento aos gráficos apresentados nas Figuras 2 e 3 apresentaremos tabelas com medidas descritivas da variável gasto, estratificadas pelas categorias ou intervalos das variáveis explicativas.

Table 3: Medidas Descritivas de Gasto com Relação a X1

Medidas	Categorias		
	Cartão de Crédito	Dinheiro	Vale Alimentação
Min.	0.10	0.09	0.06
1st Qu.	0.29	0.14	0.29
Mediana	0.48	0.21	0.48
Média	0.66	0.26	0.60
3rd Qu.	0.81	0.33	0.74
Max.	2.16	0.82	2.90
Variância	0.29	0.04	0.30
Desvio Padrão	0.54	0.19	0.55
Coeficiente de Variação	0.82	0.73	0.92

Table 4: Medidas Descritivas de Gasto com Relação a X2

Medidas	Categorias	
	Cliente Cadastrado	Cliente Não Cadastrado
Min.	0.10	0.06
1st Qu.	0.21	0.23
Mediana	0.57	0.38
Média	0.60	0.54
3rd Qu.	0.79	0.57
Max.	2.12	2.90
Variância	0.25	0.26
Desvio Padrão	0.50	0.51
Coeficiente de Variação	0.83	0.95

Table 5: Medidas Descritivas de Gasto com Relação a X3

Medidas	Faixas de Estudo em X3			
	Menor que o 1º Quartil	Entre o 1º e 2º Quartil	Entre o 2º e 3º Quartil	Maior 3º Quartil
Min.	0.06	0.10	0.09	0.10
1st Qu.	0.22	0.22	0.37	0.21
Mediana	0.32	0.33	0.52	0.46
Média	0.51	0.55	0.54	0.63
3rd Qu.	0.51	0.62	0.58	0.84
Max.	2.90	2.12	1.34	2.11
Variância	0.39	0.29	0.10	0.26
Desvio Padrão	0.63	0.54	0.32	0.51
Coeficiente de Variação	1.24	0.97	0.60	0.82

Table 6: Medidas Descritivas de Gasto com Relação a X4

Medidas	Faixas de Estudo em X4			
	Menor que o 1º Quartil	Entre o 1º e 2º Quartil	Entre o 2º e 3º Quartil	Maior 3º Quartil
Min.	0.06	0.09	0.11	0.53
1st Qu.	0.22	0.37	0.25	0.53
Mediana	0.35	0.47	0.50	0.54
Média	0.46	0.50	0.78	0.90
3rd Qu.	0.58	0.52	1.07	1.08
Max.	2.12	1.08	2.90	1.62
Variância	0.14	0.11	0.56	0.39
Desvio Padrão	0.37	0.33	0.75	0.63
Coeficiente de Variação	0.81	0.66	0.95	0.70

Com base nos gráficos e tabelas apresentadas nesta seção optou-se pela retirada da observação 82, pois esta

observação foi destacada em todos os gráficos descritos e também a sua retira altera razoavelmente as estatísticas calculadas na amostra.

2.2 Especificação do Modelo

Com base na análise descritiva será proposto um modelo linear generalizado assumindo distribuição Gama para a resposta e função de ligação inversa (canônica).

$$y_i|\underset{\sim}{x_i} \sim Gama(\theta_i, \phi_i)$$

$$\mu_i = \eta_i^{-1} = \frac{1}{\beta_0 + \beta_{11}x_{11i} + \beta_{12}x_{12i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}}$$

Abaixo definiremos as variáveis categórica inclusas no modelo, pois para estas variáveis temos categorias que são tomadas como referência.

- $X_{11} = \begin{cases} 1, & \text{se } x_1 = \text{Dinheiro} \\ 0, & \text{caso contrário} \end{cases}$
- $X_{12} = \begin{cases} 1, & \text{se } x_1 = \text{Vale Alimentação} \\ 0, & \text{caso contrário} \end{cases}$
- $X_2 = \begin{cases} 1, & \text{se } x_2 = \text{Cliente não cadastrado} \\ 0, & \text{caso contrário} \end{cases}$

2.3 Modelo Aditivo Saturado Ajustado

Após definido o modelo na seção acima, ajustamos o modelo ao conjunto de dados e foram obtidas as seguintes estimativas para os parâmetros:

Table 7: Parâmetros do Modelo

Parametro	Estimativa	Erro.Padrão
β_0	1.966	0.396
β_{11}	2.279	0.635
β_{12}	0.298	0.281
β_2	0.373	0.289
β_3	-0.005	0.047
β_4	-0.191	0.049

Com isso podemos definir nosso preditor linear, agora com as estimativas dos parâmetros.

$$\widehat{\eta_i} = 1.966 + 2.279x_{11i} + 0.298x_{12i} + 0.373x_{2i} - 0.005x_{3i} - 0.191x_{4i}$$

E na escala da variável de interesse (gasto médio) a equação é escrita:

$$\widehat{\mu_i} = \frac{1}{1.966 + 2.279x_{11i} + 0.298x_{12i} + 0.373x_{2i} - 0.005x_{3i} - 0.191x_{4i}}$$

Percebemos, pelo sentido das estimativas dos parâmetros, que para clientes que pagam em dinheiro o gasto médio estimado será menor, assim como para os clientes que optam por vale alimentação. Já para os clientes cadastrados esperamos um gasto médio maior com relação aos não cadastrados, para uma distância maior e número de pessoas elevado também espera-se um gasto médio maior.

O modelo ajustado apresentou um valor de *deviance* igual a 46.5921

2.4 Parâmetro de Dispersão ϕ

Para a distribuição Gama, associada a variável resposta, não temos o parâmetro de dispersão ϕ fixo, portanto este deverá ser estimado com base na amostra. Apresentaremos estimativas baseadas em três procedimentos de estimação diferentes.

- Baseado na estatística χ^2 de Pearson resultou em $\widehat{\phi}$ =0.5798;

- Baseado na função desvio resultou em $\widehat{\phi} = 0.501$;
- Estimativa de máxima verossimilhança,baseada na função escore resultou em $\widehat{\phi} = 0.4391$.

2.5 Modelos Alternativos

Nesta seção vamos propos alguns modelos, cujo distribuição associada é função de ligação serão as mesmas trabalhadas no modelo aditivo saturado, mas iremos alterar a combinaçao linear de parâmetros no preditor linear η. Abaixo temos os quatro preditores lineares que serão estudados, o primeiro será o aditivo estudado até aqui e os demais serão propostos agora:

- Modelo1 - Efeito de todas as variáveis explicativas.
  $\eta_i = \beta_0 + \beta_{11}x_{11i} + \beta_{12}x_{12i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$;
- Modelo2 - Apenas o efeito de X1, X2 e X4.
  $\eta_i = \beta_0 + \beta_{11}x_{11i} + \beta_{12}x_{12i} + \beta_2 x_{2i} + \beta_4 x_{4i}$;
- Modelo3 - Apenas o efeito de X2 e X4.
  $\eta_i = \beta_0 + \beta_2 x_{2i} + \beta_4 x_{4i}$;
- Modelo4 - Efeito de X2 e X4 considerando a interação entre elas.
  $\eta_i = \beta_0 + \beta_2 x_{2i} + \beta_4 x_{4i} + \beta_5 x_{4i} x_{2i}$.

Abaixo temos uma tabela com medidas de ajuste para cada um dos modelos.

Table 8: Medidas de Ajuste para os Modelos Propostos

Modelo	Nparameters	LogLikMax	Deviance	X2Pearson	PseudoR2	AIC
1	6	-17.8599	46.5987	53.8356	0.4013	49.7198
2	5	-17.8647	46.6028	53.5603	0.4012	47.7294
3	3	-30.3681	58.6801	64.1607	0.2460	68.7362
4	4	-30.0821	58.3743	65.3195	0.2500	70.1641

Percebemos através da Tabela 8 que há uma semelhança entre os modelos 1 e 2 e entre os modelos 3 e 4 e também é nítida a diferença entre essas duas duplas. A primeira dupla de modelos (modelos 1 e 2) apresentaram um poder de explicação bem maior do que os modelos 3 e 4, devido ao possível efeito siginificativo das variável consideradasnestes modelos. Dentre os modelos 1 e 2 percebemos que há uma boa semelhança em quase todas as medidas de ajuste, indicando que o efeito da variável X3 pode não ser significativo, note que o critério de Akaike é menor para o modelo 2, pois esta medida penaliza os modelos pelo numéro de parâmetros.

Para comprovar os indícios observados na tabela 8 faremos uma sequência de testes estatísticos para comparação de modelos, os testes a seguir serão baseados na razão de verosimilhanças. Note nos testes abaixo que a distribuição adotada para a estatística do teste será a F de Snedecor, pois o parâmetro de dispersão precisou ser estimado.

- Modelo1 vs Modelo2

  Hipóteses
  $\begin{cases} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{cases}$

```
## Analysis of Deviance Table
##
## Model 1: Gasto ~ X1 + X2 + X4
## Model 2: Gasto ~ X1 + X2 + X3 + X4
##   Resid. Df Resid. Dev Df Deviance    F Pr(>F)
## 1        94       46.6
## 2        93       46.6  1  0.00752 0.01   0.91
```

  Como o p-valor foi extremamente alto (> 0.9) não rejeitamos a hipótese nula, ou seja, o efeito da variável X3 não é siginificamente importância para explicar a variável resposta, confirmando os indícios observados anteriormente.

- Modelo1 vs Modelo3

  Hipóteses
  $\begin{cases} H_0 : \beta_{11} = \beta_{12} = \beta_3 = 0 \\ H_a : \beta_{1i} \neq 0 \text{ e/ou } \beta_3 \neq 0 \end{cases}$

```
## Analysis of Deviance Table
##
## Model 1: Gasto ~ X2 + X4
## Model 2: Gasto ~ X1 + X2 + X3 + X4
##   Resid. Df Resid. Dev Df Deviance    F  Pr(>F)
## 1        96       58.0
## 2        93       46.6  3     11.4 6.56 0.00045 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  Como o p-valor foi extremamente baixo (< 0.001) temos evidências para rejeitar a hipótese nula, ou seja, o efeito das variáveis X1 e X3, conjuntamento, não pode ser desprezado no modelo, pois são significamente importantes para explicar a variável resposta.

- Modelo3 vs Modelo4

  Hipóteses
  $\begin{cases} H_0 : \beta_5 = 0 \\ H_a : \beta_5 \neq 0 \end{cases}$

```
## Analysis of Deviance Table
##
## Model 1: Gasto ~ X2 + X4
## Model 2: Gasto ~ X2 * X4
##   Resid. Df Resid. Dev Df Deviance    F Pr(>F)
## 1        96         58
## 2        95         58  1   0.0399 0.06   0.81
```

  Como o p-valor foi alto (> 0.80) não temos evidências para rejeitar a hipótese nula, ou seja, o efeito referente a interação entre X2 e X4 não é significativamente importante no modelo.

Perceba que não é correta a comparação entre os modelos 1 e 4 pelo teste de razão de verossimilhanças, pois eles não são modelos encaixados, isto é, não há uma restrição de parâmetros que os torne equivalentes.

2.6 Testes de Hipóteses

Considerando o modelo 1, apresentaremos nesta subseção alguns testes de hipóteses para os parâmetros do modelo.

No software estatístico R temos duas função equivalentes que realizam análise de variância de modelos, são elas as funções *anova* e *car::Anova* que retornam valores particulares. Aplicaremos as duas funções no modelo 1 e apresentaremos seus resultados.

```
anova(model1, test = "F")

## Analysis of Deviance Table
##
## Model: Gamma, link: inverse
##
## Response: Gasto
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev     F  Pr(>F)
## NULL                    98       66.2
## X1    2    11.49        96       54.7  9.91 0.00013 ***
## X2    1     0.75        95       53.9  1.29 0.25844
```

```
## X3    1     0.17        94       53.8  0.30 0.58607
## X4    1     7.17        93       46.6 12.37 0.00068 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(model1, test = "F")

## Analysis of Deviance Table (Type II tests)
##
## Response: Gasto
## Error estimate based on Pearson residuals
##
##             SS Df     F  Pr(>F)
## X1        11.4  2  9.84 0.00013 ***
## X2         0.9  1  1.48 0.22752
## X3         0.0  1  0.01 0.90956
## X4         7.2  1 12.37 0.00068 ***
## Residuals 53.9 93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que as estatísticas do testes (apresentadas na coluna F) e suas respectivas significâncias são distintas entre os testes, isto se dá pois a função *anova* realiza testes sequenciais $(\beta_0; \beta_1|\beta_0; \beta_2|\beta_1, \beta_0; ...)$, ou seja, leva em consideração a ordem que as variáveis entraram no modelo. Já a função *Anova*, da bilioteca *car*, faz os teste os efeitos considerando todas as variáveis no modelo $(\beta_1|\beta_0, \beta_2, ...\beta_p; \beta_2|\beta_0, \beta_1, ...\beta_p; ...)$, ou seja, não é importante a ordem de entrada das variáveis no modelo. Normalmente o interesse está em testar os efeitos das variáveis com todas as demais já no modelo, portanto a função *Anova* é mais indicada.

Uma outra alternativa para testar o efeito dos parâmetros no modelo é considerando o teste de Wald, que utiliza a distribuição assintótica dos estimadores de máxima verossimilhança.

```
summary(model1)

##
## Call:
## glm(formula = Gasto ~ X1 + X2 + X3 + X4, family = Gamma(link = "inverse"),
##     data = da)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.433  -0.607  -0.197   0.234   1.924
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.96582    0.39626    4.96  3.2e-06 ***
## X1Dinheiro                2.27857    0.63549    3.59  0.00054 ***
## X1Vale Alimentação        0.29754    0.28069    1.06  0.29187
## X2Cliente não cadastrado  0.37294    0.28944    1.29  0.20078
## X3                       -0.00537    0.04723   -0.11  0.90965
## X4                       -0.19075    0.04900   -3.89  0.00019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.5798)
##
##     Null deviance: 66.178  on 98  degrees of freedom
## Residual deviance: 46.592  on 93  degrees of freedom
## AIC: 45.32
##
## Number of Fisher Scoring iterations: 6
```

As significâncias deste teste agora estão associadas a cada parâmetro presente no modelo, ou seja, para a variável X1 como temos três categórias que a definem teremos dois parâmetros no modelo e serão apresentadas siginificância para estes dois parâmetros. Na figura acima podemos observar que as variáveis X3 e X2 não acrecentam grande poder de explicação no modelo, devido as seus altos p-valores, ainda é importânte ressaltar

que para o parâmetro β_{12} (representado por X1Vale Alimentação) apresentou um alto p-valor, porém por estar associado a variável X1, e um de seus dois parâmetros apresentou alta significância, não é pertinente interpretar a variável como de efeito não significativo.

Com as estimativas para os parâmetros em mãos podemos construir intervalos de confiança e novamente temos duas metodologias que serão apresentadas: a primeira será baseada no perfilamento da verossimilhanças (LogLik) e a segunda será baseada na estatística de Wald.

Table 9: Intervalos de Confiança para os Parâmetros (95% de confiança)

Parâmetros	LogLik		Wald	
	2.5%	97.5%	2.5%	97.5%
β_0	1.210417	2.765588	1.189159	2.742483
β_{11}	1.147657	3.648989	1.033037	3.524112
β_{12}	-0.238562	0.874893	-0.252600	0.847685
β_2	-0.251507	0.895279	-0.194361	0.940240
β_3	-0.099438	0.085835	-0.097953	0.087203
β_4	-0.280506	-0.088778	-0.286795	-0.094713

Observamos pela Tabela 9 que há diferenças entre os intervalos de confiança baseados no perfil de verossimilhança e baseados na estaística de Wald, esta diferença se dá pois o teste de Wald se baseia na normalidade assintótica dos estimadores de máxima verossimilhança. Perceba que mesmo com as diferenças pontuais dos intervalos, não houve divergências nas interpretações, fazendo ligação com os testes de hipóteses, interpreta-se como efeitos não significativos aqueles nos quais o valor zero está contido no intervalo.

2.7 Seleção de Variáveis

Nesta seção utilizaremos do algoritmo *stepwise*, que fará a permutação de variáveis dentro do modelo, com o critério de seleção de variáveis critério de Akaike (AIC), pois este penaliza os modelos com um número excessivo de parâmetros.

Primeiramente faremos a permutação de variáveis a serem inclusas no modelos utilizando como modelo completo o modelo aditivo com X1, X2, X3 e X4. E posteriormente consideraremos como modelo completo o modelo considerando todas as variáveis explicativas X1, X2, X3 e X4 e mais suas interações duplas. Abaixo são apresentadas a última iteração do algoritmo para ambas as especificações.

- Considerando como modelo completo o aditivo com X1, X2, X3 e X4.

```
## Step:  AIC=43.27
## Gasto ~ X1 + X4
##
##        Df Deviance    AIC
## <none>      47.450 43.266
## + X2    1   46.600 43.764
## + X3    1   47.448 45.262
## - X4    1   54.684 54.048
## - X1    2   58.274 58.391
##
## Call:  glm(formula = Gasto ~ X1 + X4, family = Gamma(link = "inverse"),
##     data = da)
##
## Coefficients:
##        (Intercept)          X1Dinheiro  X1Vale Alimentação                  X4
##             2.2233              2.2262              0.2787            -0.1866
##
## Degrees of Freedom: 98 Total (i.e. Null);  95 Residual
## Null Deviance:      66.18
## Residual Deviance: 47.45  AIC: 43.27
```

- Considerando como modelo completo o modelo com X1, X2, X3, X4 e mais suas interações de segunda ordem.

```
## Step:  AIC=43.27
## Gasto ~ X1 + X4
##
##         Df Deviance    AIC
## <none>       47.450 43.266
## + X2     1   46.600 43.764
## + X3     1   47.448 45.262
## + X1:X4  2   46.642 45.839
## - X4     1   54.684 54.048
## - X1     2   58.274 58.391
##
## Call:  glm(formula = Gasto ~ X1 + X4, family = Gamma(link = "inverse"),
##     data = da)
##
## Coefficients:
##        (Intercept)          X1Dinheiro  X1Vale Alimentação                  X4
##             2.2233              2.2262              0.2787             -0.1866
##
## Degrees of Freedom: 98 Total (i.e. Null);  95 Residual
## Null Deviance:      66.18
## Residual Deviance: 47.45  AIC: 43.27
```

Em ambas as especificações o algoritimo, utilizando o AIC como critério de seleção, nos retornou o mesmo modelo, ou seja, não há interações de segunda ordem que sejam relevantes para o modelo assim como as variável X2 e X3. Observe que utlizando o algoritmo, chegamos no modelo que também seria encontrado utlizando as análises anteriores, pois as variáveis X2 e X3 foram as que apresentaram fortes indícios de não significância para o modelo.

Portanto como modelo proposto ajustado temos:

$$
\begin{aligned}
y_i|\underset{\sim}{x_i} &\sim Gama(\theta_i, \phi_i) \\
\widehat{\mu_i} &= \frac{1}{2.223 + 2.226x_{11i} + 0.279x_{12i} - 0.187x_{4i}}
\end{aligned}
\tag{1}
$$

3 Aplicação do modelo

Como exemplo didático continuaremos com o modelo 1 e o utilizaremos para estimador o gasto médio de clientes com os seguintes perfis:

Table 10: Perfil de indíviduos para estimação

Indíviduo	X2	X1	X3	X4
1	Cliente cadastrado	Dinheiro	5.0	2.0
2	Cliente cadastrado	Dinheiro	5.0	5.0
3	Cliente cadastrado	Cartão de crédito	5.0	2.0
4	Cliente cadastrado	Vale Alimentação	5.0	2.0

Com base na tabela 10 foram estimados os gastos médios para cada indivíduo, o erro padrão da estimativa e seus respectivos intervalos de confiança. Abaixo temos uma tabela com essas medidas.

Table 11: Estimativas para o Gasto Médio e Intervalo de Confiança

Indivíduo	Estimativas		Intervalo de Confiança	
	Gasto Estimado	Erro Padrão	Lower 2.5%	Upper 97.5%
1	0.261	0.044	0.175	0.347
2	0.306	0.061	0.188	0.425
3	0.642	0.120	0.407	0.877
4	0.539	0.102	0.339	0.739

Então para o primeiro indivíduo, um cliente cadastrado que pagou sua última compra em dinheiro, reside a 5 km do mercado e tem 2 pessoas morando wm sua casa, estima-se um gasto médio de 175 a 347 reais, da mesma forma para os outros indivíduos. Note que a única diferença entre o primeiro e segundo indivíduo é o aumento no número de pessoas que moram com ele e perceba que a estimativa para o gasto médio também aumentou, já entre os indivíduos 3 e 4 a diferença está na forma de pagamento e temos para o indivíduo que optou pela forma de pagamento vale alimentação um gasto médio estimado menor. Este acréscimo e decréscimo na estimativa do gasto médio com relação as variáveis explicativas já era esperado, veja a interpretação na seção 2.3.