UNIVERSIDADE DE ÉVORA

2º Relatório de Regressão e Classificação

Regressão multinomial e análise discriminante

Trabalho realizado no âmbito da cadeira de Regressão e Classificação por

Jorge Couto	58656
Miguel Grilo	58387

Colégio Luís António Verney

Introdução

No âmbito da consolidação das novas matérias lecionadas em aula após a realização do primeiro trabalho, foi-nos pedida a realização de um segundo, e último, projeto, usando a análise classificatória, discriminante e multinomial.

O objetivo do trabalho assentou na classificação dos indivíduos observados na nossa base de dados com recurso à análise classificatória e, usando as técnicas de modelagem da análise discriminante e multinomial, tentar criar um modelo que permitisse prever precisamente novos indivíduos corretamente com base nas suas características.

Seguindo o trabalho anterior, optamos pela reutilização da mesma base de dados Diamonds, disponibilizada pelo pacote ggplot2. Essa base de dados contra com várias variáveis, apesar de termo-nos focado no uso das variáveis:

- price – O preço em dólares americanos;
- carat – O peso do diamante (sem medida de peso referida na fonte oficial);
- cut – A qualidade do diamante, dividida pelas categorias *Fair, Good, Very Good, Premium* e *Ideal* (uma variável categórica ordinal);
- depth – A porcentagem de profundidade total, dada pela fórmula z / mean(x, y);
- table – A largura do topo do diamante relativamente ao ponto mais largo (novamente, sem medida referida na fonte oficial).

Exercício 1 – Análise Classificatória

Para começo, foi-nos requisitado o uso da análise classificatória para dividirmos as amostras observadas em pelo menos 3 grupos distintos. Foi-nos, também, pedido pelo uso de pelo menos 2 métodos diferentes para comparação dos grupos formados.

Antes de tudo, utilizamos o comando *set.seed* para que trabalhássemos sempre sobre os mesmos dados, de modo a não se notarem variações significativas nos nossos *outputs*. De seguida, criamos um subconjunto com somente 1000 amostras da base de dados original, para facilitar no código, e as variáveis a serem usadas, descartando aquelas que não seriam utilizadas para a criação dos grupos. Apesar de já verificado no trabalho anterior, decidimos, por protocolo, estudar novamente a existência de dados omissos que foi, mais uma vez, negada.

De seguida, categorizamos a variável resposta price em quatro categorias, com base nos quartis: Baixo, do valor de preço menor até 0.25; Médio-baixo, de 0.25 até 0.5; Médio-alto, de 0.5 a 0.75 e Alto, de 0.75 ao valor de preço maior. Fizemos isto para poder melhor perceber as diferenças entre grupos no quesito gráfico. Também, tornámos a nossa variável categórica em numérica. Finalmente, criamos um subset apenas com as variáveis explicativas, para que a variável resposta não influenciasse a formação dos grupos, visto que é com base nessa que queremos dividir os grupos.

Com todas as preparações prontas, decidimos seguir para a análise classificatória a usar o método simples e completo, com uma matriz das distâncias euclidianas em ambos os casos. Apesar de sempre criarmos dendogramas, fizêmo-lo apenas para ter uma visão geral da unificação dos grupos. Com 1000 amostras, os dendogramas são praticamente ilegíveis, então utilizamos o *scree plot* como recurso para a decisão do ponto de corte.

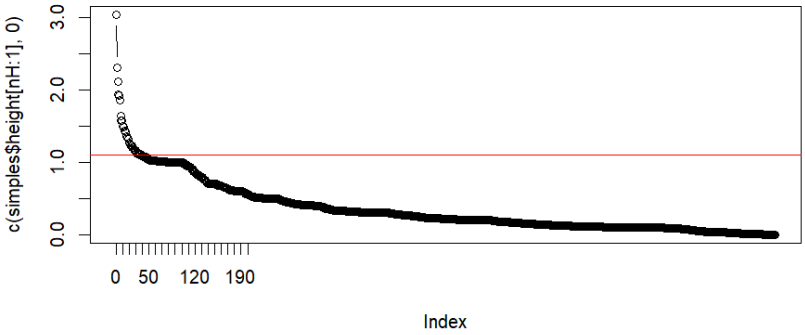

Fig. 1 – Screeplot, com ponto de corte, do método simples.

Todavia, cortando com ponto de corte fixo ou com número de grupos fixos, o resultado foi o mesmo, seguindo pelo método simples ou completo. Acabamos por obter vários grupos singulares, com apenas ou ou dois indivíduos, e apenas um ou dois grupos globais a envolver todas as outras amostras. Portanto, decidimos usar outros métodos para tentar procurar grupos mais equilibrados, visto que um grupo com um só indivíduo acaba por não ajudar na classificação de outras amostras.

Depois de estandartizarmos as variáveis, tentamos novamente utilizar o método simples e completo com matriz das distâncias euclidianas. No entanto, os resultados foram iguais, se não mais desequilibrados ainda. Sem outra escolha, recorremos ao método Ward.D2, o mais correto no quesito da obtenção de grupos mistos, continuando ainda a usar a matriz das distâncias euclidianas. Recorremos ao Ward.D2 utilizando os dados normais e estandartizados.

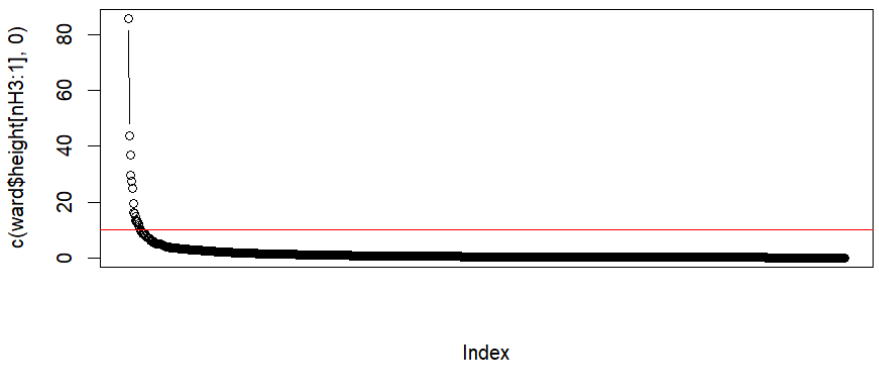

Fig. 2 – Screeplot, com ponto de corte, do Ward.D2 com dados normais.

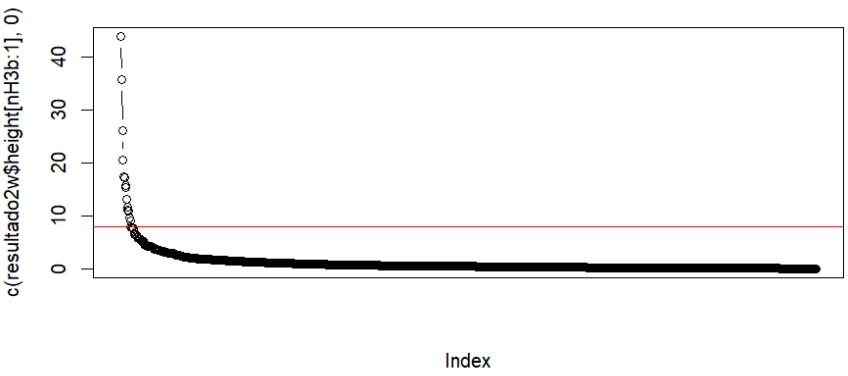

Fig. 3 – Screeplot, com ponto de corte, do Ward.D2 com dados estandartizados.

Usando as variáveis normais, obtivemos 18 grupos, com o menor grupo tendo apenas 6 amostras. Por outro lado, usando as variáveis estandartizadas conseguimos uma melhor divisão, tendo 15 grupos, e com o menor grupo tendo 12 amostras. Portanto, optamos por utilizar os grupos obtidos pelo Ward.D2 com variáveis estandartizadas.

Entretanto, obter os grupos por si só não chega. Queremos, também, reduzir a quantidade de grupos que temos, de modo a facilitar a análise. Reobtivemos o *price* como variável contínua ao criar a variável dadosprice composta pelos mesmos 1000 indivíduos escolhidos apenas com a variável de preço original (contínua), e com essa utilizamos o teste de games howell para verificar a significância entre grupos. O objetivo é verificar se existem diferenças significativas entre diferentes grupos e, se não, uni-los. No entanto, não podemos unir grupos ao calhas, precisando de unir grupos gradualmente: Vendo através da ordem das médias do preço, devemos garantir, para unir um grupo, que todos os outros grupos entre esses dois são, também, significativos com os dois a se unir, de modo a garantir que a união dos grupos faz sentido. Assim, obtivemos somente cinco grupos:

- O grupo formado pelos grupos 11, 10 e 1 (11_10_1);
- O grupo formado por 5, 8, 4, 6 e 12 (5_8_4_6_12);
- O grupo formado por 14, 3, 9 e 15 (14_3_9_15);
- O grupo formado por 7 e 13 (7_13);
- O grupo 2, individual.

É, com esses grupos, que seguimos para os exercícios seguintes.

Exercício 2 – Análise Discriminante

Para a análise discriminante, criamos primeiro a variável nvalido que representa o número de amostras total a serem usadas, neste caso 1000. De seguida, dividimos os dados entre dados para ajuste do modelo e dados para teste, com 75% das amostras para um lado e 25% para outro. Estranhamente, acabou por não ser uma divisão perfeita seguindo a regra dos

0.25/0.75, ficando 725 amostras para ajuste e 275 para teste. Algo pouco importante, mas que achámos pertinente mencionar.

Através da matriz de correlações garantimos que nenhuma das variáveis está altamente correlacionada entre si. Se estivessem, poderíamos reduzir a quantidade de variáveis total no modelo ao incluir apenas uma delas. Neste caso, não houve nenhum par de variáveis com correlação acima dos 0.8.

Em seguida, criamos um subset dos dados somente com as variáveis explicativas dos dados de ajuste com base nos grupos unificados. Esse mesmo subset foi usado para o método *stepwise* com o critério lambda de Wilks. De acordo com o método, todas as variáveis são significativas, o que significa que todas entrarão no modelo discriminante como variáveis explicativas. Utilizando a análise discriminante, obtemos 3 funções, sendo a primeira a que melhor explica a variância dos resultados, e sendo *carat* a variável que mais discrimina nessa mesma função. Nas outras, muito menos explicativas, o *cut* é a variável mais discriminativa.

Com o modelo criado, fez-se importante validar os pressupostos. De nada valeria o modelo se fosse incerto, afinal. O pressuposto da inexistência de multicolinearidade, visto anteriormente, pode ser validado pela falta de correlações acima dos 0.8 entre as variáveis explicativas. O pressuposto da normalidade, no entanto, verificado tanto por recurso ao *QQPlot* quanto ao teste de *Mardia*, não pode ser validado. No entanto, a análise discriminante é robusta à violação do pressuposto da normalidade multivariada contando que a dimensão da amostra menor dos grupos seja superior ao número de variáveis discriminantes, o que é o caso tanto para os dados de ajuste quanto de teste, e que as médias dos grupos não sejam proporcionais às suas variâncias. Infelizmente, isso não se pode verificar, pois o grupo 2 apresenta todas as suas amostras na mesma categoria da variável *cut* (total separação dos dados), o que torna a sua variância nula e, portanto, o cálculo da proporção infinita. Depois de questões à professora quanto ao método correto de procedimento nesta situação, concluímos que o método multinomial seria mais correto do que a análise discriminante neste caso. Foi-nos sugerido continuar com a análise discriminante, mas não tentar a análise discriminante quadrática por não valer a pena. Então, fizemos como sugerido.

Avaliando a função discriminante linear através da matriz de confusão, obtivemos um modelo altamente significativo, com uma precisão de 74.6% e uma taxa de acerto ao acaso de apenas 44.4%, tornando o modelo muito significativo (valor p inferior a 0.001). A especificidade geral é bastante alta, mas a sensibilidade varia entre muito alta e muito baixa, sendo alta (acima dos 0.8) para os grupos unificados 14_3_9_15, 5_8_4_6_12 e 7_13 mas baixa (0.26 e 0.37) para os grupos 2 e 11_10_1. Verificando o modelo do ponto de vista da reação a novas observações, isso é, usando agora os dados de teste para a matriz de confusão, obtemos um modelo menos preciso, mas ainda significativo, com uma taxa de acerto de 75.8% e uma taxa de acerto ao calhas de 40.5%, tornando-o também muito significativo (valor p inferior a 0.001). A mesma especificidade alta global e verifica, e agora os extremos de sensibilidade tornam-se mais óbvios, especialmente no grupo 2, onde a sensibilidade torna-se de apenas 0.29, mas a especificidade alcança os 100%.

Exercício 3 – Análise Multinomial

Antes de começarmos a análise multinomial de facto, decidimos melhor representar os grupos com nomes que melhor mostram os seus preços. Através de vários gráficos, conseguimos perceber como os preços, e as variáveis explicativas, variam entre cada um dos grupos.

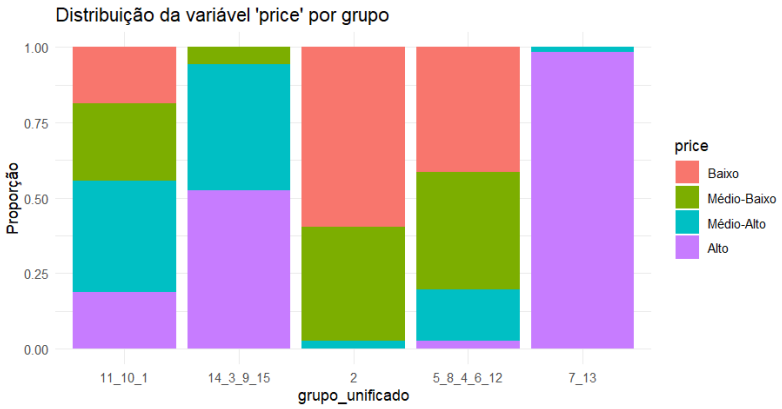

Fig. 4 – Distribuição do preço categorizado por cada um dos grupos.

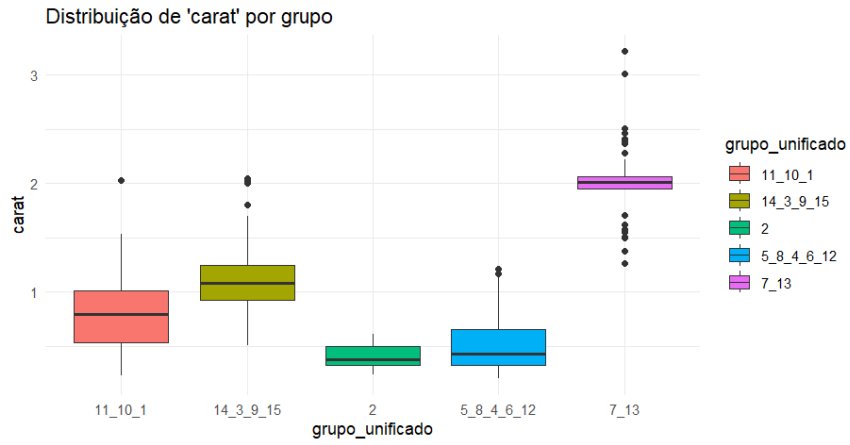

Fig. 5 – Distribuição do carat por cada um dos grupos.

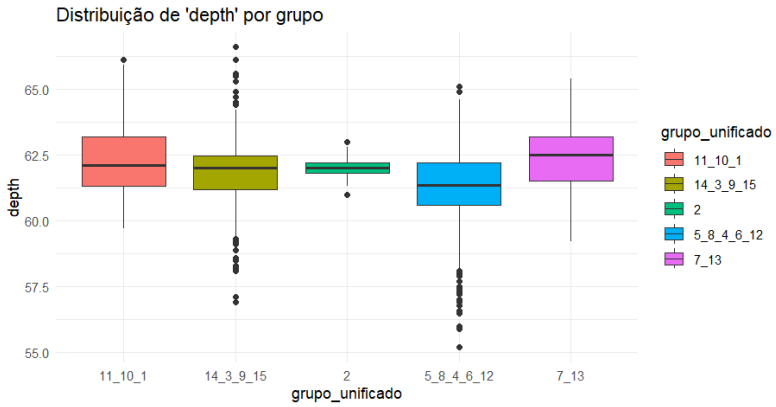

Fig. 6 – Distribuição de depth por cada um dos grupos.

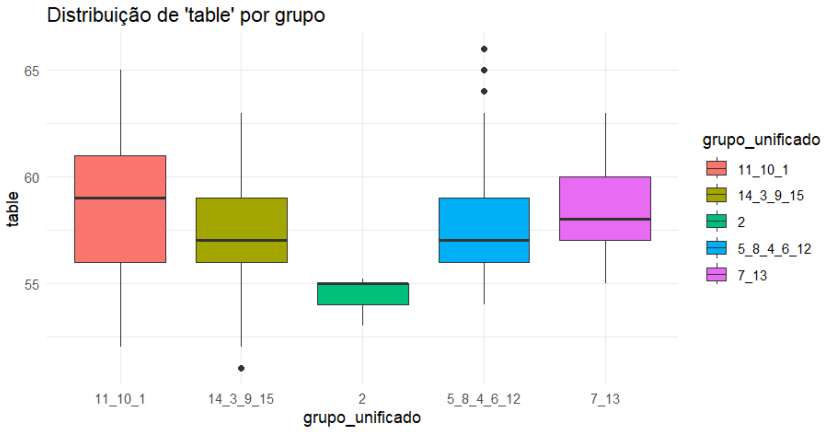

Fig. 7 – Distribuição de table por cada um dos grupos.

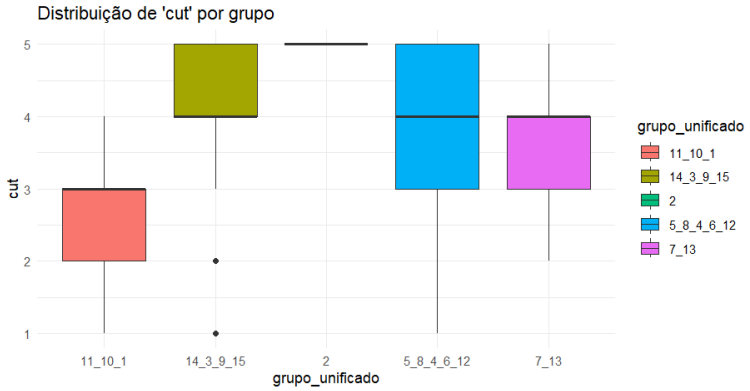

Fig. 8 – Distribuição das categorias de cut por grupo.

Através da análise gráfica, chegamos aos seguintes nomes para cada um dos grupos:

- Equilibrados – Originalmente o grupo 11_10_1, tem diamantes em todas as categorias de preço, um depth geralmente elevado e carat baixo a intermédio, com table e cut ambos intermédios;
- Intermédios-Altos – Originalmente o grupo 14_3_9_15, tem diamantes entre o preço médio-alto e alto. Têm todas as variáveis explicativas intermédias, salvo pelo cut que é, geralmente, elevado;
- Baratos – Originalmente o grupo 2, tem mais diamantes de preço baixo, apesar de ter alguns de preço médio-baixo e poucos de preço médio-alto. Têm o carat e table mais baixos e a menor variedade em depth. Todos os seus diamantes têm o cut mais alto;
- Intermédios-Baixos – Originalmente o grupo 5_8_4_6_12, tem diamantes em todas as categorias de preço, mas com mais no preço baixo e médio-baixo. Têm o segundo menor carat, table mediano e tanto o cut quanto o depth são altamente variáveis;
- Luxo – Originalmente o grupo 7_13, tem os diamantes mais caros, com quase todos no preço alto, salvo poucas exceções no preço médio-alto. Têm o maior carat, embora com vários outliers, depth e table equilibrados e um cut variável, mas geralmente bom.

Seguindo agora para a análise multinomial de facto, começamos por usar o *crosstable* para verificar se todas as variáveis são, ou não, significativas, algo que conseguimos validar, com os valores p associados a todas inferiores a 0.001. Em seguida, ajustamos os modelos univariados, com base nos grupos anteriormente mencionados, colocando o grupo dos equilibrados como referência. Para qualquer uma das variáveis, os valores p retornados foram menores que 0.20, o que nos permitiu começar o modelo multinomial múltiplo com todas as variáveis. De seguida, passamos para o passo seguinte, tentando simplificar o modelo a verificar se podemos remover as variáveis não significativas do modelo em ordem decrescente do valor p do teste TRV. Pensando assim, a primeira variável que tentamos remover foi o cut. Todavia, como a diferença entre o modelo multivariado com e sem cut é muito significativa (valor p < 0.001), não podemos remover a variável cut sem tirar capacidade explicativa ao modelo. O mesmo aconteceu com as seguintes variáveis que tentamos remover (depth e table), ficando assim com o modelo completo ainda.

Apesar de termos escrito o código para adicionar interações no modelo multinomial, acabámos por comentar esse código, não o incluindo no nosso modelo final. O motivo por detrás dessa decisão assenta na complexidade do nosso modelo final que, com todas as interações que seriam adicionadas, tornar-se-ia demasiado complexo de analisar. Portanto, o nosso modelo final é aquele que usa todas as variáveis explicativas para tentar concluir em qual dos grupos unificados o indivíduo cai.

Com o modelo criado, seguimos então para a validação dos pressupostos. Começando pela multicolinearidade, tentamos usar o *vif*. No entanto, sendo incompatível com multinom ou mlogit, criamos o modelo usando glm para verificar o vif. De acordo com o teste, os valores vif de cada uma das variáveis são todos inferiores a 5, o que nos permite validar este pressuposto por não haver multicolinearidade significativa. As variáveis são estatisticamente independentes. O pressuposto da linearidade não pode ser validado, visto que de acordo com o teste todas as transformações aplicadas são não-lineares e significativas. No entanto, aplicando as transformações sugeridas e excluindo as variáveis colineares, obtemos um modelo menos explicativo do que o nosso modelo original, visto que apresenta um AIC maior e um R quadrado menor do que o nosso modelo original. Decidimos investigar sobre, e acabamos por concluir que, para o tipo de modelo que procuramos, poderíamos seguir com o modelo que temos, sem transformações, todavia não-linear. Portanto, seguimos assim.

Avaliando a adequabilidade do modelo e a bondade do ajustamento, verificamos uma *Residual Deviance* de 1091.452 e um AIC igual a 1131.452. Vendo agora os testes, rejeitamos a hipótese de um bom ajuste com g = 10 no teste *Hosmer & Lemeshow* com um valor p inferior a 0.001. Também rejeitamos a hipótese com g = 6, tendo um valor p muito significativo (< 0.001). Verificando os R quadrados, obtemos um R quadrado de McFadden igual a 0.596, aproximadamente 0.6, e de Nagelkerke igual a 0.386, aproximadamente 0.39. Apesar de não muito altos, os valores já demonstram alguma capacidade discriminativa pelo modelo.

Analisando, agora, os resíduos com o teste de *Cessie* e o coeficiente de *Brier*, verificamos que apenas o grupo dos diamantes intermédios-altos têm um valor p acima do nosso alfa de 0.05, indicando um bom ajuste para esse grupo. O nosso coeficiente de Brier é 0.215, um pouco superior a 0.2.

Finalmente, analisamos a matriz de confusão para verificar a capacidade discriminativa do nosso modelo. O nosso modelo multinomial apresenta uma precisão de 0.8, bastante superior à chance de acerto sem informação de apenas 0.47. O modelo é, então, bastante significativo, com um valor p inferior a 0.001. A sensibilidade geral do modelo é alta, exceto para o grupo dos diamantes equilibrados, em que a sensibilidade é de 0.5119, pouco superior a 0.5. A especificidade, por outro lado, não sofre esse problema, sendo superior a 0.9 para todos os grupos. Ou seja, o modelo têm um bom desempenho em todos os grupos exceto a classificar corretamente diamantes que sejam de facto do grupo dos equilibrados.

Em termos pontuais, e utilizando o grupo dos diamantes equilibrados de referência, podemos concluir que quanto maior o peso do diamante maior a chance de ser de luxo ou intermédio-alto comparado aos diamantes equilibrados, e menor a chance de ser barato ou intermédio-baixo comparado aos equilibrados. Comparado às outras variáveis explicativas, o peso têm a maior grandeza na equação, indicando assim que é um forte preditor do nível de preço. Quanto ao depth, quanto maior o depth maior a chance de ser um diamante de luxo, barato ou intermédio-alto e menor a chance de ser intermédio-baixo comparado aos equilibrados. Podemos dizer, portanto, que a profundidade está relacionada com preços mais extremos. O table é o exato oposto, tendo um coeficiente negativo para todos os grupos. Quer isso dizer que diamantes com maior table têm maior chance de ser equilibrados comparados a outros. Por fim, cut têm um coeficiente positivo para todos os grupos, indicando que cortes melhores aumentam as chances do diamante ser de um grupo específico. Isto torna-se especialmente notável para os diamantes baratos, com o coeficiente do cut para esses sendo significativamente superior aos outros. Ou seja, o corte influencia significativamente a chance do diamante ser barato comparado aos diamantes equilibrados. Através do cálculo dos intervalos de confiança, podemos obter o intervalo a 95% de cada um dos coeficientes que participam nas equações:

Intervalos de Confiança

	Intercept	Carat	Depth	Table	Cut
Intermédios-Altos	[-2.756, 9.902]	[4.521, 7.002]	[-0.039, 0.191]	[-0.427, -0.208]	[1.216, 1.796]
Baratos	[21.429, 22.067]	[-23.418, -13.327]	[0.279, 1.991]	[-5.155, -3.394]	[29.469, 32.661]
Intermédios-Baixos	[21.530, 42.211]	[-4.874, -3.071]	[-0.511, -0.269]	[-0.197, -0.002]	[0.288, 0.735]
Luxo	[-67.155, -65.623]	[11.444, 16.142]	[0.668, 1.017]	[-0.378, 0.041]	[0.877, 1.961]

Exercício 4 – Comparação dos Modelos

Uma vez com os modelos discriminantes e multinomial criados, falta agora saber qual modelo é o mais adequado para predição de sujeitos novos. Apesar de já termos anteriormente concluído que a análise multinomial seria mais adequada que a análise discriminante pela

violação dos pressupostos, os problemas quanto à linearidade do nosso modelo multinomial podem, ainda assim, tornar a análise discriminante a melhor preditora.

Através das matrizes confusão, verificamos que, apesar de terem uma precisão mais ou menos semelhante, a perca da sensibilidade da análise discriminante comparada à multinomial é bastante significativa, havendo dois grupos na análise discriminante com sensibilidade péssima, enquanto na multinomial somente o grupo de referência (os diamantes equilibrados) sofre de baixa sensibilidade (que não é tão baixa quando comparada à análise discriminante, inclusive). Também, apesar de termos tido problemas com os pressupostos em ambos os modelos, a violação dos pressupostos da análise discriminante é muito mais grave do que aquela vista na análise multinomial.

Portanto, apesar da precisão bastante significativa dos dois modelos, torna-se claro que o modelo multinomial é, neste caso, o mais adequado entre os dois modelos para predizer o grupo de diferentes indivíduos.

Exercício 5 – Predição para Indivíduos ao Acaso

Com os dois modelos criados, podemos tentar predizer qual seria o grupo mais provável para um diamante com certas características. Para o teste realizado, assumimos um diamante com um *carat* de 0.75, um *depth* de 61, um *table* de 58 e um *cut* muito bom (Very Good).

Através da análise discriminante, um diamante que apresente essas características tem 69% de chance de fazer parte do grupo dos diamantes intermédios-baixos, sendo a segunda melhor possibilidade a chance de fazer parte do grupo dos diamantes equilibrados, com 21% de chance. Utilizando a análise multinomial, chegamos à mesma conclusão. A chance de um diamante com essas características pertencer ao grupo dos diamantes intermédios-baixos é de 65%, e a segunda melhor possibilidade é que faça parte do grupo dos diamantes equilibrados, com 28% de chance. Portanto, podemos concluir que, neste caso, a previsão dos dois modelos converge, e que um diamante com as características anteriormente mencionadas provavelmente fará parte do grupo dos diamantes intermédios-baixos.