Regressão logística na identificação de factores de risco em acidentes automóveis e fraude de seguros.

José Luís Mourão

Faculdade de Ciências
Universidade do Porto

28 de Janeiro de 2013

Modelo de Regressão Logística

- A regressão logística é útil para prever uma variável binária dependente de variáveis preditoras.
- O modelo de regressão logística é dado por $\pi(\mathbf{x}) = P(Y = 1|\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}}$.
- $g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- Note-se que não há uma relação linear entre a variável dependente e o conjunto das variáveis independentes.

Modelo de Regressao Logística

- Aplica-se a transformação *logit*: $ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_i X_i$, onde:
  - β_0 é a constante do modelo;
  - β_i são os coeficientes das variáveis independentes;
  - X_i é o conjunto de variáveis independentes (contínuas ou discretas).

- Esta transformação é importante visto possuir muitas das propriedades de um modelo de regressão linear.

Modelo de Regressão Logística

- Aplicando a exponencial à equação anterior, obtém-se: $\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0} e^{\beta_i X_i}$.

- Pode-se constatar que quando o valor de uma variável independente aumenta em 1 unidade e todas as outras variáveis se mantêm constantes, o novo rácio é dado por: $\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0} e^{\beta_i X_i} e^{\beta_k}$

- O factor e^{β_k} é chamado de *odds ratio* (OR) e varia entre 0 e ∞. Indica a quantidade relativa que a probabilidade do resultado ser 1 aumenta (OR > 1) ou diminui (OR < 1).

Ajustamento do modelo

- Em regressão logística, não existe um valor de R^2 como no caso da regressão linear.
- Como alternativa, usa-se um pseudo R^2:
  - Cox & Snell Pseudo-R^2;
  - Nagelkerke Pseudo-R^2.
- Ou o teste de Hosmer-Lemeshow.

Case Study 1

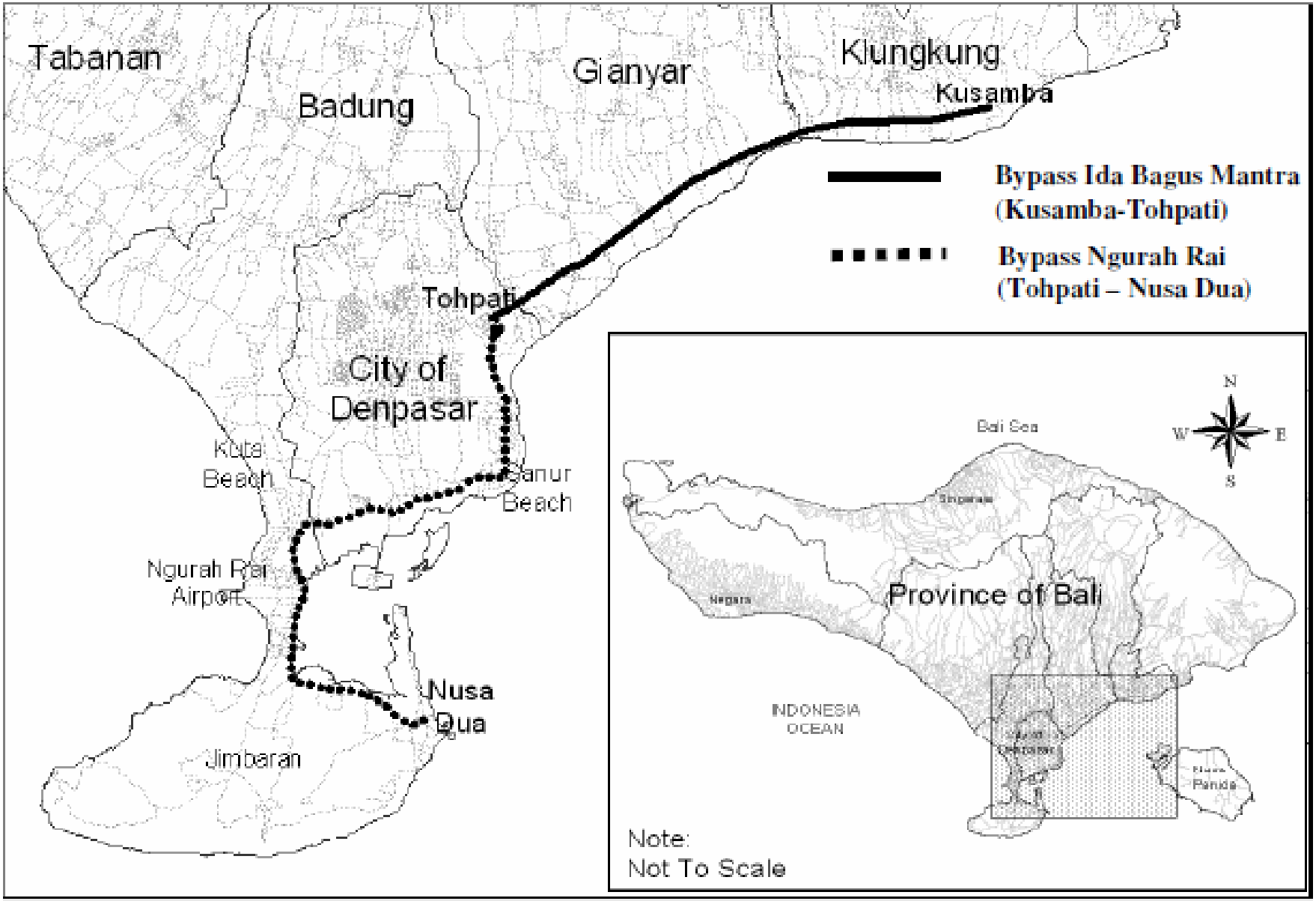

Figure 1 Case study area

Case Study 1

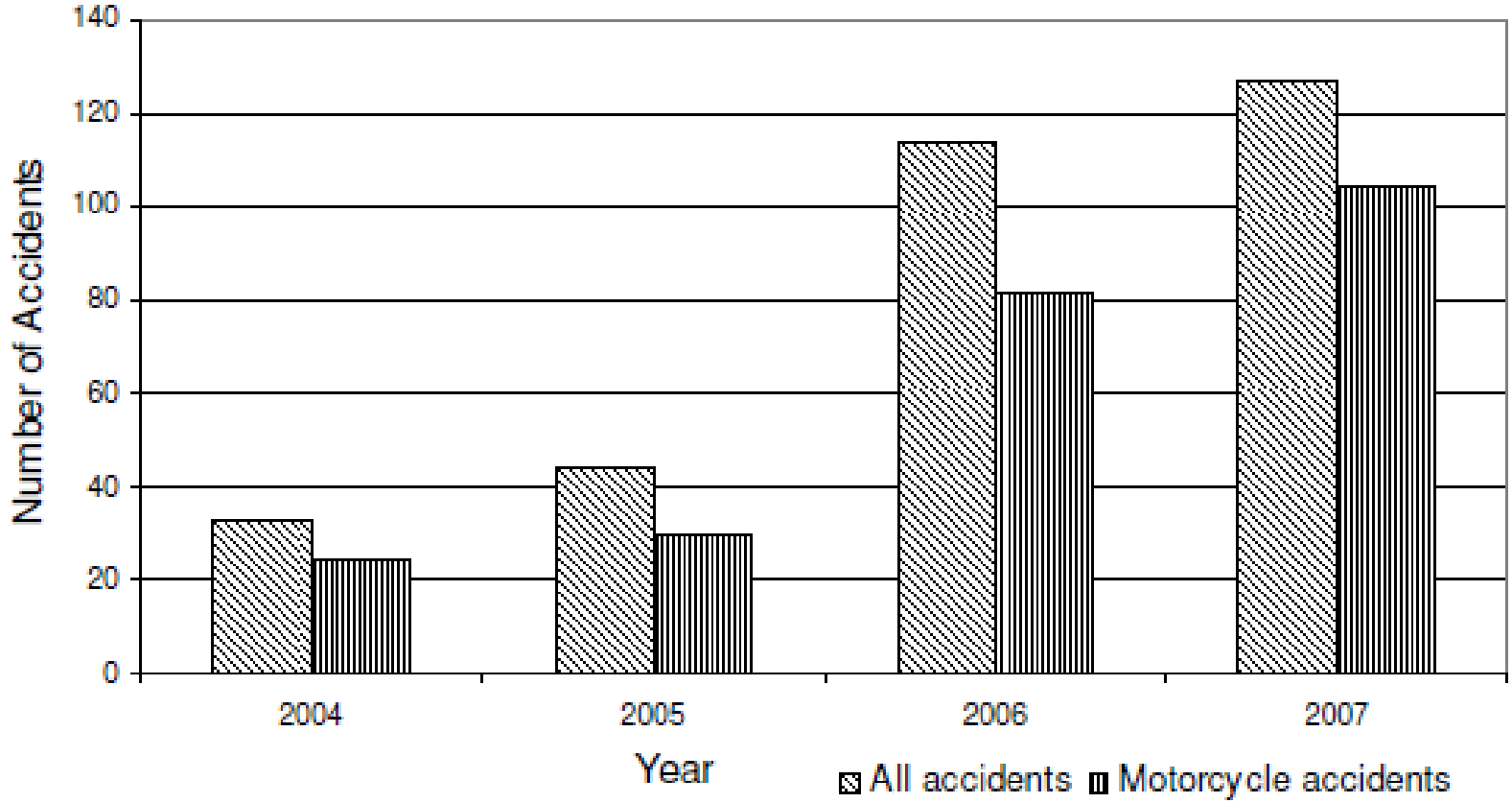

Figure 2 Road accidents based on modes of transport

Case Study 1 - Desenvolvimento do Modelo

Table 1 Variables selected for the study

No.	Variable Type	Classifications and Coding	Variable Title
1.	Fatal Accidents	0 = Non Fatal Accidents 1 = Fatal Accidents	Fatal
2.	Accident type	0 = with fixed object 1 = overturned 2 = with vehicles	Atyp
3.	Collision type	0 = Out of Control 1 = Right Angle, 2 = Side Swipe 3 = Rear End 4 = Head On	Ctyp
4.	Vehicle type (at fault)	0 = Heavy vehicle 1 = Light vehicle 2 = Motorcycle	Veh
5.	Accident cause	0 = Others 1 = Speeding 2 = Run red light 3 = Follow too close 4 = Wrong way 5 = Failed to yield	Caus
6.	Accident Location	1 = Link 0 = Junction	Loc
7.	Time of accident	1 = Day time 0 = Night time	Time
8.	Gender (of driver/motorcyclist at fault)	1 = Male 0 = Female	Gender
9.	Age (of driver/ motorcyclists at fault)	Year	Age

Case Study 1 - Desenvolvimento do Modelo

- Algumas das classificações das variáveis podem ser neglicenciadas dada a sua baixa proporção.

- Teste de hipóteses para proporções.
  - $H_0 : p = 0$
  - $H_1 : p \neq 0$

Case Study 1 - Desenvolvimento do Modelo

Table 2 Hypothesis testing: data statistics (motorcycle fatal accidents)

Description	X	N	P-value	95% Confidence level	
				Lower	Upper
Accident Type					
With fixed object *	9	240	0.038	0.0	0.1
Overturned	31	240	0.129	0.1	0.2
With vehicles	200	240	0.833	0.9	0.9
Collision Type					
Out of control	45	240	0.188	0.1	0.2
Right angle	51	240	0.213	0.2	0.3
Side swipe	24	240	0.100	0.1	0.1
Rear end	63	240	0.263	0.2	0.3
Head on	57	240	0.238	0.2	0.3
Vehicle Type (at fault)					
Heavy vehicle	21	240	0.088	0.1	0.1
Light vehicle	56	240	0.233	0.2	0.3
Motorcycle	163	240	0.679	0.6	0.7
Accidents Cause					
Others	52	240	0.217	0.2	0.3
Speeding*	7	240	0.029	0.0	0.1
Run red light*	3	240	0.013	0.0	0.0
Follow too close	60	240	0.250	0.2	0.3
Wrong way	67	240	0.279	0.2	0.3
Failed to yield	51	240	0.213	0.2	0.3
Gender (of driver at fault)					
Male	210	240	0.875	0.8	0.9
Female	30	240	0.125	0.1	0.2
Location					
Link	202	240	0.842	0.8	0.9
Junction	38	240	0.158	0.1	0.2
Time of Accident					
Day time	121	240	0.504	0.4	0.6
Night time	119	240	0.496	0.4	0.6

* Statistically insignificant at the 5% level; the 95% confidence limits include 0.

Case Study 1 - Desenvolvimento do Modelo

- 3 factores podem ser excluídos/alterados:
  - Acidente com objecto imóvel - excluído
  - Acidente devido a velocidade excessiva
  - Acidente devido a passagem com sinal vermelho

  } agregados

- O mesmo se verifica na análise relativa aos acidentes de veículos motorizados.

Case Study 1 - Análise do Modelo

Table 4 Goodness of fit (pseudo R^2 and H-L test)

Pseudo R^2 Test			
Fatal Accidents Model	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Motorcycle	295.666	.142	.190
Motor vehicle	412.368	.083	.111
Hosmer and Lemeshow Test (H-L Test)			
	Chi-square	df	Sig.
Motorcycle	1.627	8	.990
Motor vehicle	3.957	8	.861

- A medida de Nagelkerke indica que o 1º modelo explica 19% da variância da variável dependente e o 2º modelo explica 11% da variância da variável dependente.
- O teste de Hosmer-Lemeshow mostra que ambos os modelos são significativos (valor p > 0.05).

Case Study 1 - Resultados

Table 6 Model results

	Motorcycle fatal accidents			Motor vehicle fatal accidents		
Variables	B	Sig.	Exp(B)	B	Sig.	Exp(B)
Location(1)	.427	.300	1.532	.210	.557	1.234
Atyp(1)	-.150	.872	.861	-.051	.924	.951
Atyp(2)	**-1.525***	.195	.218	.176	.805	1.193
Ctyp(1)	-.697	.458	.498	**-1.067***	.154	.344
Ctyp(2)	-.549	.569	.578	**-1.195***	.137	.303
Ctyp(3)	-.497	.592	.608	-.898	.218	.407
Ctyp(4)	-.348	.702	.706	-.511	.504	.600
Time(1)	.073	.802	1.076	.036	.883	1.037
Cause(1)	**-1.261***	.175	.283	.256	.680	1.291
Cause(2)	-.493	.565	.611	.442	.486	1.555
Cause(3)	-.963	.235	.382	-.014	.983	.986
Cause(4)	**-1.600****	.063	.202	-.618	.331	.539
Age	.002	.877	1.002	**.019***	.136	1.019
Gender(1)	**-1.331*****	.006	.264	**-.952*****	.015	.386
Veh(1)	.110	.845	1.116	.162	.734	1.176
Veh(2)	-.382	.462	.683	.144	.748	1.155
Constant	**3.604*****	.018	36.727	.331	.744	1.392

Bold figures are significant as follows:
*** Significant at 95%, ** Significant at 90% , * Significant at 80%.

Case Study 1 - Alguns Resultados

- O elevado coeficiente da constante no modelo de acidentes fatais de motociclos indica que existem outros factores (infraestrutura da estrada ou condições da superfície da estrada, por exemplo), que podem ter impacto nas fatalidades.
- No modelo relativo a todos os veículos motorizados, os tipos de colisão "ângulo recto"e "colisão lateral"estão negativamente relacionados com acidentes fatais, sendo 30% menos prováveis a causar fatalidades do que acidentes devidos a falta de controlo.
- A idade do motorista influencia positivamente a ocorrência de fatalidades.
- O condutor ser mulher é menos provável a influenciar acidentes fatais nos 2 modelos do que ser homem, sendo que a probabilidade de um acidente ser fatal devido a um condutor do sexo feminino é 30% e 40%, respectivamente, mais baixa do que para um condutor do sexo masculino.

Case Study 2 - Tomada de decisões baseada no custo

- Não é razoável assumir custos de classificação errada iguais, visto que não é isso que se verifica em muitas situações da vida real.
- Exemplo: falhar na detecção de uma doença pode ter consequências fatais mas uma predição positiva falsa será, em princípio, menos sério.
- No caso em questão, podemos supôr que o caso de fraude é relativamente raro, estando, no entanto, associado a um custo elevado caso não seja detectado.

Case Study 2 - Tomada de decisões baseada no custo

- Assim, vamos minimizar

$$\arg\min_{t \in 0,1} \sum_{j=0}^{1} p(j|x) C_{t,j}(x) \tag{1}$$

- Matriz de custos

	Situação honesta	Situação burla
Previsão honesta	$C_{0,0}(x)$	$C_{0,1}(x)$
Previsão burla	$C_{1,0}(x)$	$C_{1,1}(x)$

Case Study 2 - Tomada de decisões baseada no custo

- Assume-se que a matriz de custos respeita 2 condições:
  - Nenhuma linha domina outra;
  - O custo de uma classificação incorrecta é sempre superior do que o custo de uma classificação correcta.

- Pode-se verificar que o critério (1) se traduz na seguinte regra:

$$p(j=1|x) > \frac{C_{1,0}(x) - C_{0,0}(x)}{C_{1,0}(x) - C_{0,0}(x) + C_{0,1}(x) - C_{1,1}(x)} \rightarrow 1 \quad (2)$$

Case Study 2 - Tomada de decisões baseada no custo

- A matriz de custos pode ser vista da seguinte forma:

	Situação honesta	Situação burla
Previsão honesta	0	Montante da indemnização
Previsão burla	Custo da auditoria	Custo da auditoria - Montante da indemnização

Case Study 2 - Conjunto de Dados

- Amostra aleatória de pedidos de cobertura de acidentes de carros em Espanha durante o ano 2000.
- n = 2403, com 2229 pedidos honestos e 174 fraudulentos (92.76% e 7.24%, respectivamente).
- A proporção de pedidos fraudulentos está dentro do intervalo mais citado e reportado pela indústria dos seguros, 5-10%.

Case Study 2 - Conjunto de Dados

Name	Type	Explanation
Fraud	Nominal	Observed type of claim (fraudulent equals 1, legitimate 0)
Cov1	Nominal	Third party liability equals 1, otherwise 0
Cov2	Nominal	Third party liability plus arson/theft/glass breakage equals 1, otherwise 0
Veh1	Nominal	Automobile for private use equals 1, otherwise 0
Veh2	Nominal	Motorcycle equals 1, otherwise 0
Age	Continuous	Age of insured driver when the accident occurred
Gender	Nominal	Insured driver is male equals 1, female 0
Records	Continuous	Number of previous claims of the insured
Policyage	Continuous	Number of years the insured has been with the company
Fault	Nominal	The other driver accepts fault for the accident equals 1, otherwise 0
NFS	Nominal	Use of the no-fault system equals 1, otherwise 0
Weekend	Nominal	Accident occurring on a weekend equals 1, otherwise 0
Delay	Nominal	Claim not reported to the company within the legally established period equals 1, otherwise 0
Claim amount	Continuous	Insurer's valuation of the vehicle damages (once the deductible has been discounted, if it exists)
Audit cost	Continuous	Cost of fraud investigation

Case Study 2 - Conjunto de Dados

- Utilizando as médias das variáveis *Claim amount* e *Audit Cost* na amostra para preencher a matriz de custos, obtém-se:

	Situação honesta	Situação burla
Previsão honesta	0 €	818.14 €
Previsão burla	72.26 €	-745.88 €

- Suponha-se o caso de ter um portefólio de 500.000 apólices; neste caso, se não se investisse no controlo da burla, a companhia perderia aproximadamente 2.468.328 €. E bastaria detectarem-se 70% de todas as burlas para uma poupança de 769.486 €.

Case Study 2 - Resultados

- Foram estudados 6 cenários que diferem entre si na informação relativa aos custos:
  - Caso 1: não há qualquer informação.
  - Caso 2: todos os custos individuais são conhecidos.
  - Caso 3: são utilizados custos médios.
  - Caso 4: são utilizados custos médios.
  - Caso 5: os custos das indemnizações são conhecidos para cada elemento do conjunto de dados mas os custos das auditorias são calculados utilizando a informação disponível.
  - Caso 6: os custos das indemnizações são conhecidos para cada elemento do conjunto de dados mas os custos das auditorias são calculados utilizando a informação disponível.

Case Study 2 - Resultados

Table 9
Estimation results

	Scenario 1a		Scenario 1b		Scenario 2		Scenario 3	
	Coefficients	*P* value	Coefficients	*P* value	Coefficients	*P* value	Coefficients	*P* value
Constant	−3.733	0.000[a]	−3.497	0.000[a]	42.156	0.000[a]	−3.733	0.000[a]
Cov1	0.271	0.512	0.299	0.472	−0.388	0.850	0.271	0.512
Cov2	0.547	0.061[b]	0.529	0.074[b]	0.210	0.908	0.547	0.061[b]
Vehuse 1	0.388	0.281	0.387	0.287	−1.540	0.362	0.388	0.281
Vehuse 2	1.526	0.001[a]	1.495	0.002[a]	2.401	0.306	1.526	0.001[a]
Age	0.002	0.754	0.004	0.604	0.021	0.565	0.002	0.754
Gender	0.396	0.095[b]	0.324	0.175	0.533	0.674	0.396	0.095[b]
Records	0.085	0.144	0.081	0.168	−0.099	0.747	0.085	0.144
Policyage	−0.030	0.097[b]	−0.028	0.120	−0.030	0.722	−0.030	0.097[b]
Fault	−0.548	0.392	−0.527	0.374	−2.584	0.623	−0.548	0.392
NFS	0.174	0.782	0.187	0.747	2.248	0.667	0.174	0.782
Weekend	0.208	0.255	0.121	0.516	−0.864	0.410	0.208	0.255
Delay	0.218	0.194	0.372	0.031[a]	1.860	0.086[b]	0.218	0.194
*Claim amount**			0.523	0.000[a]	−7.206	0.000[a]		
*Audit cost**					23.713	0.000[a]		
	Dependent variable: *Fraud* Ss = 2403; LL = −604.23; LL_0 = −624.36 LR test = 40.28 (*P* value = 0.000)		Dependent variable: *Fraud* Ss = 2403; LL = −583.80; LL_0 = −624.36 LR test = 81.13 (*P* value = 0.000)		Dependent variable: *Fraud* Ss = 2403; LL = −28.20; LL_0 = −624.36 LR test = 1192.38 (*P* value = 0.000)		Dependent variable: *Fraud* Ss = 2403; LL = −604.23; LL_0 = −624.36 LR test = 40.28 (*P* value = 0.000)	
Predictive accuracy								
True Negative	66.44%		67.52%		99.42%		18.35%	
False Negative	45.98%		37.90%		0.57%		10.34%	
False Positive	33.56%		32.48%		0.58%		81.65%	
True Positive	54.02%		62.10%		99.43%		89.66%	
Summary cost information (€)								
True Negative	0.00		0.00		0.00		0.00	
False Negative	107,574.19		34,653.03		1,856.71		12,812.54	
False Positive	44,771.44		55,058.70		3,131.83		109,204.75	
True Positive	−81,296.42		−146,405.07		−168,019.94		−160,550.77	
Total cost	71,049.21		−56,693.34		−163,031.40		−38,533.48	
Average cost per claim	29.57		−23.59		−67.85		−16.04	

Case Study 2 - Resultados

Table 9 (*continued*)

	Scenario 4		Scenario 5		Scenario 6			
	Coefficients	*P* value	Coefficients	*P* value	Coefficients	*P* value	Coefficients	*P* value
Constant	−3.497	0.000[a]	−2.818	0.000[a]	−2.888	0.000[a]	−1.509	0.000[a]
Cov 1	0.299	0.472	0.156	0.000[a]	0.144	0.000[a]	0.124	0.121
Cov 2	0.529	0.074[b]	0.138	0.000[a]	0.115	0.000[a]	0.087	0.144
Vehuse 1	0.387	0.287	0.020	0.565	0.000	0.990	−0.063	0.332
Vehuse 2	1.495	0.002[a]	0.014	0.787	−0.115	0.003[a]	−0.084	0.315
Age	0.004	0.604	0.001	0.441	0.000	0.606	0.002	0.063[b]
Gender	0.324	0.175	0.055	0.027[a]	0.032	0.094[b]	−0.000	0.994
Records	0.081	0.168	−0.001	0.912	−0.007	0.158	0.027	0.016[a]
Policyage	−0.028	0.120	−0.002	0.234	−0.000	0.955	−0.002	0.472
Fault	−0.527	0.374	0.016	0.808	0.061	0.229	−0.215	0.072[b]
NFS	0.187	0.747	0.013	0.843	0.005	0.916	0.143	0.203
Weekend	0.121	0.516	0.022	0.335	0.009	0.608	−0.009	0.780
Delay	0.372	0.031[a]	0.034	0.102	0.005	0.776	−0.025	0.392
*Claim amount**	0.523	0.000[a]	0.321	0.000[a]	0.263	0.000[a]	0.480	0.000[a]
*Audit cost**								
	Dependent variable: *Fraud* Ss = 2403; LL = −583.80; LL_0 = −624.36 LR test = 81.13 (*P* value = 0.000)		Dependent variable: *Audit cost** Ss = 2403; R^2 = 31.87% $F_{(13,2389)}$ = 85.96 (*P* value = 0.000)		Dependent variable: *Audit cost** (NF) Ss = 2229; R^2 = 36.62% $F_{(13,2215)}$ = 98.46 (*P* value = 0.000)		Dependent variable: *Audit cost** (F) Ss = 174; R^2 = 84.58% $F_{(13,160)}$ = 67.51 (*P* value = 0.000)	
Predictive accuracy								
True Negative	62.31%		57.16%				59.47%	
False Negative	33.33%		28.74%				29.31%	
False Positive	37.69%		42.84%				40.53%	
True Positive	66.67%		71.26%				70.69%	
Summary cost information (€)								
True Negative	0.00		0.00				0.00	
False Negative	21,999.75		18,845.18				19,466.63	
False Positive	62,990.15		69,203.59				67,419.03	
True Positive	−156,822.11		−158,708.17				−158,281.93	
Total cost	−71,832.20		−70,659.41				−71,396.27	
Average cost per claim	−29.89		−29.40				−29.71	

Case Study 2 - Cenário 1: Classificação insensível aos custos

- Neste caso, trata-se de um classificador baseado no erro.
- O tipo de veículo, género do cliente, a cobertura da apólice e o número de anos que o cliente está com a companhia parecem ser estatisticamente significantes relativamente à probabilidade de burla.
- Corresponde a um custo total de 210.079 €.
- Uma examinação mais detalhada revelou que este modelo classifica todos os casos como honestos, logo não é útil.

Case Study 2 - Cenário 2: Todos os custos são conhecidos

- Este cenário assume que a seguradora tem acesso a todos os custos - montante de indemnização e custo da auditoria.
- Este cenário corresponde a um custo de -163.031 €.
- Não é um caso prático já que normalmente só se conhece o custo da auditoria no final do processo e, por regra, está directamente relacionado com a presença de burla.

Case Study 2 - Cenário 3: Custos médios

- São utilizados os valores anteriormente referidos para os custos.
- Este modelo apresenta resultados maus, com uma percentagem de 23.51% casos correctamente classificados.
- Corresponde a um custo total de -38.533 €.

Case Study 2 - Cenário 4: Custos individuais e médios

- A seguradora tem acesso à informação de cada montante de indemnização individual e ao custo médio de auditoria.
- A percentagem de casos correctamente classificados é de 62.63%.
- Corresponde a um custo total de -71.832 €.

Case Study 2 - Cenário 5: Custos individuais e custos previstos

- Em vez de se simplesmente assumir que o custo da auditoria é o custo médio das auditorias, esse custo vai ser previsto utilizado um modelo de regressão linear baseado nas mesmas variáveis utilizadas até agora.
- $R^2 = 31.87\%$, logo não há um bom ajustamento do modelo aos dados, não sendo aconselhável o uso desta técnica.
- A percentagem de casos correctamente classificados é de 58.18%.
- Corresponde a um custo total de -70.659 €, ligeiramente pior do que o caso pior.
- Uma alternativa seria "melhorar" a predição do custo de auditoria.

Case Study 2 - Cenário 6: Custos individuais e custos previstos

- Semelhante ao caso anterior mas usam-se 2 modelos de regressão linear: um para o subconjunto dos casos de burla e outro para o subconjunto dos casos honestos.
- Parte do pressuposto que os custos de auditoria são diferentes para cada uma das situações.
- $eac(x) = \hat{p}(x) \times acf(x) + (1 - \hat{p}(x)) \times ach(x)$.
- Depois do custo esperado para uma auditoria ser calculado, essa informação é combinada com o custo da indemnização e utiliza-se o modelo do caso 4 para calcular a probabilidade de burla.
- A percentagem de casos correctamente classificados é de 58.59%, com um custo de -71.396 €, o que não é uma melhoria significativa em relação ao caso anterior.

Case Study 2 - Conclusões

- A decisão baseada em custos mostra um claro progresso em relação à decisão baseada em erros.
- No entanto, tendo em conta o cenário ideal (2) e o cenário mais perto da realidade, ainda se verifica uma discrepância com um valor financeiro de 91.199 €.
- Tendo em conta os estudos realizados nos cenários 5 e 6, poderia pensar-se em utilizar alternativas (não-lineares, por exemplo) que melhorassem a predição do custo das auditorias.