Proval Parcial 1 - Amanda Freitas Carnaiba - NUSP 13485660

Amanda Freitas Carnaiba

October 22, 2022

1 Parte 1

Considere a seguinte tabela com os resultados de quatro modelos distintos de regressão. A variável dependente é uma variável Y não especificada aqui.

	(1)	(2)	(3)	(4)
X_1	**a,b0**	**a,(b+1)2**	**(a+2),(b+3)1**	**a,(b+1)4**
	(0,02)***	(0,01)***	(a,b0)	(0,02)***
X_2		-4,12	**f,gh**	-3,85
		(1,90)*	(f,g+1)	(1,02)***
X_3			-0,15	
			(0,20)	
X_4				2,45
				(0,25)***
Constante	2,10	1,54	-3,29	1,29
	(c,de)	(0,56)**	(2,81)*	(0,01)***
N	150	150	150	150
R^2	31%	35%	87%	41%
Erros padrão robustos em parêntesis				
*** p < 0,1%; ** p < 1%; * p < 5%				

Figure 1: Tabela 1 – Resultados da Regressão – Variável Dependente: Y

Note que a tabela está preenchida com letras em alguns lugares ao invés de algarismos. Para converter estas letras em números, é preciso tomar o seu número USP e proceder da forma a seguir (aqueles que não possuem número USP podem usar o RG):

Considere o seguinte número USP: 12345678

Associamos para cada letra um dos numerais em ordem: a = 1; b = 2; c = 3; e assim sucessivamente até h = 8. Nos termos da tabela em que constam as letras, você deve substituir pelo algarismo correspondente. Como exemplo, para a expressão da primeira linha do modelo 2 [a, (b+1)2], o termo deve ser substituído por: 1,(2+1)2 = 1,32.

0) Assim, antes de iniciar, você deve reescrever a tabela acima com todos os algarismos completados. Note que é preciso utilizar 8 algarismos em 6 valores diferentes.

	(1)	(2)	(3)	(4)
X1	1,30	1,42	3,61	1,44
	(0,02)***	(0,01)***	(1,30)	(0,02)***
X2		-4,12	6,60	-3,85
		(1,90)*	(6,70)	(1,02)***
X3			-0,15	
			(0,20)	
X4				2,45
				(0,25)***
Constante	2,10	1,54	-3,29	1,29
	(4,85)	(0,56)**	(2,81)*	(0,01)***
N	150	150	150	150
R2	31%	35%	87%	41%
Erros padrão robustos em parêntesis *** p < 0,1%; ** p < 1%; * p < 5%				

Figure 2: Tabela 2 – Resultados da Regressão – Variável Dependente: Y - NUSP 13485660

A partir do que é apresentado nesta tabela, responda às questões a seguir:

1) Escreva a equação que representa o único modelo bivariado estimado. Interprete-a:

$$Y = 2.10 + (1.30 \times X_1) + \varepsilon$$

Neste modelo, quando X_1 é igual a 0, Y é igual a 2.10. A cada aumento de 1 unidade na variável independente X_1, a variável dependente Y aumenta em 1.30 unidades.

2) O que significam os asteriscos ao lado do termo em parênteses? Apresente o teste de hipóteses que justifica esta indicação. Interprete o resultado obtido para o coeficiente estimado de X_1 para o primeiro modelo;

Os asteriscos dizem respeito à significância estatística. No primeiro modelo, com um erro padrão de 0.02, temos um p-valor menor que 0.01%. Isto significa que podemos rejeitar a hipótese nula de que X_1 não está associada à variável dependente Y.

Podemos realizar um teste de hipótese para confirmar isso, em que:

$$H_0 : \beta_{X1} = 0$$
$$H_1 : \beta_{X1} \neq 0$$

Tendo que $\beta_{X1} = 1.30$ e $se(\beta_{X1}) = 0.02$, podemos calcular o intervalo de confiança à 99,7% de significância estatística:

Parâmetro de X1	Desvio Padrão de X1	IC (99,7%)	
1.3	0.02	1.24	1.36

Como 0 não está contido no Intervalo de Confiança, podemos rejeitar a hipótese nula de que $\beta_{X1} = 0$.

3) Explique também o que significam os termos apresentados em parênteses em toda a tabela. Explique em que situação se deve apresentar os valores calculados desta forma;

Os termos em parênteses ao longo de toda tabela são os erros-padrão dos termos de cada variável independente X, com os asteriscos indicando a significância estatística a partir do p-valor. Por exemplo, vemos que, no terceiro modelo, nenhuma variável independente X tem significância estatística, isto é, o p-valor de todas elas está acima de 5%, o que nos leva a não rejeitar a hipótese nula de que seus parâmetros são iguais a 0, consequentemente concluindo que essas variáveis não explicam a variação na variável dependente Y. Além disso, é importante apresentar os erros-padrão para o cálculo do intervalo de confiança dos termos estimados.

4) Corrija o nível de significância do parâmetro estimado para o termo da constante do primeiro modelo, mostrando como chegou ao resultado;

O termo da constante do primeiro modelo não possui significância estatística, como podemos comprovar por meio de um teste de hipóteses usando o Intervalo de Confiança:

Constante	Desvio Padrão	IC (99,7%)		IC (95%)		IC (68%)	
2.1	4.85	-12.45	16.65	-7.6	11.8	-2.75	6.95

Como o valor 0 está contido em todos os intervalos, podemos concluir que a constante não possuí nível de significância estatística.

5) Escreva a equação para o 2º modelo apresentado. Interprete cada um dos parâmetros estimados;

$$Y = 1.54 + (1.42 \times X_1) + (-4.12 \times X_2) + \varepsilon$$

Neste modelo, quando X_1 e X_2 são iguais a 0, Y é igual a 1.54. Com o aumento de uma unidade em X_1, controlado por X_2, isto é, mantendo X_2 constante, a variável Y aumenta 1.42 unidades. Já quando ocorre o aumento de uma unidade em X_2 controlando por X_1, isto é, mantendo X_1 constante, ocorre a diminuição de 4.12 unidades em Y.

6) Como é possível comparar este resultado com o obtido no primeiro modelo? Apresente um teste que permita dizer se o novo parâmetro se alterou em relação ao primeiro modelo. Discuta;

Podemos testar a hipótese de que o parâmetro de X_1 mudou com a introdução de X_2 a partir do intervalo de confiança. No modelo 1, temos um resultado $\beta_{X1} = 1.30$ e $se(\beta_{X1}) = 0.02$. Já no modelo 2, temos mudança nos parâmetros de X_1: $\beta_{X1} = 1.42$ e $se(\beta_{X1}) = 0.01$. Com isso temos o seguinte intervalo de confiança, para 95% de significância estatística:

Parâmetro de X1	Desvio Padrão de X1	IC (95%)	
1.30	0.02	1.26	1.34
1.42	0.01	1.40	1.44

Com isso, é possível verificar que o valor do parâmetro realmente se alterou do modelo 1 para o modelo 2, pois os valores, considerando o intervalo de confiança, não se interceptam.

7) O que significa o termo R^2? Interprete o valor encontrado para o segundo modelo;

O termo R^2 indica a fração de variação da variável dependente Y que é explicado pelas variáveis independentes do modelo. Está sempre entre 0 e 1, isto é, entre 0% e 100%, e esperamos que introduzir novas variáveis explicativas aumentem o valor de R^2, isto é, a fração de variação da variável dependente que queremos que nosso modelo seja capaz de explicar. Com a introdução de X2 no segundo modelo, o R^2 passa de 31% para 35%, o que significa que as variáveis X1 e X2 explicam 35% da variação da variável dependente Y.

No 3º modelo indicado na tabela, há dois problemas. O primeiro é uma violação clássica de uma das hipóteses do modelo de MQO; o segundo parece ser um erro de digitação a respeito da significância de um dos parâmetros estimados.

8) Explique qual é a violação mencionada e quais são as evidências que suportam a sua interpretação. Quais são os impactos desta violação na interpretação dos resultados?

No terceiro modelo, temos uma alteração importante no coeficiente de X2: no segundo modelo, tínhamos $\beta_{X2} = -4.12$ e $se(\beta_{X2}) = 1.90$, com significância estatística ao nível de 95%. Já com a introdução da variável X3, o parâmetro de X2 mudou de sinal e o erro-padrão também aumento, sendo que agora temos $\beta_{X2} = 6,60$ e $se(\beta_{X2}) = 6,70$. Isso indica que pode haver problema de multicolinearidade com as variáveis independentes, o que viola as hipóteses necessárias para o modelo MQO. Um R^2 muito alto, porém sem significância estatística no modelo, pode ser indicativo de multicolinearidade, assim como um erro-padrão muito alto. Se X2 e X3 são altamente correlacionadas, é praticamente impossível fazer apontamentos sobre de que forma cada uma dessas variáveis está relacionada à variável Y. Além disso, a baixa significância estatística nos levaria a conclusões de que nenhuma dessas variáveis explicam a variável dependente Y, quando na verdade elas explicam. Podemos identificar a multicolinearidade usando o fator de inflação da variância (VIF), ou uma matriz de correlação entre as variáveis. Caso seja detectada a multicolinearidade, a solução é remover uma dessas duas variáveis do modelo.

9) Como podemos identificar este erro de digitação? Discuta.

No terceiro modelo, o valor da Constante é -3.29, com um erro-padrão de 2.81. A tabela indica que há significância estatística a 95% para esta estimativa, porém há um claro erro de digitação: utilizando intervalo de confiança, à 95% de significância estatística temos que o valor do termo Constante desse modelo está entre -8.91 e 2.33. Dado que o valor 0 está contido nesse intervalo, não podemos rejeitar a hipótese nula de que não há relação com a variável Y, isto é, de que o valor da Constante pode ser igual a 0, portanto não há significância estatística e não deveria haver um asterisco ao lado deste valor.

10) Quais efeitos produzem a introdução de uma nova variável explicativa em um modelo de regressão? Discuta utilizando os resultados obtidos no 4º modelo em comparação com os demais.

No quarto modelo, a variável X3 foi removida, provavelmente por causar problemas de multicolinearidade, e houve a introdução da variável X4. Em relação ao modelo 2, houve um aumento no valor de R^2, que passou de 35% para 41%, o que significa que o modelo 4 tem a capacidade de explicar 41% da variação da variável dependente Y. Todas as variáveis no quarto modelo são estatisticamente significantes ao nível de 99,9%, o que é outro indicativo de que este é um bom modelo. Para melhorar este modelo, poderia ser introduzidas mais variáveis independentes que aumentassem o valor de R^2 para algum valor acima de 50%, porém sem perder a significância estatística das demais variáveis. Em termos de interpretação dos coeficientes, um modelo multivariado indica unidades de variação numa variável dependente que são causadas por variação em uma variável independente, controlada pelas demais. Isto é, por exemplo, a variação de uma unidade em X4 causa a variação de 2,45 unidades em Y, controlada pelas variáveis X1 e X2, isto é, pressupondo que as variáveis X1 e X2 não variam. No entanto, a introdução de novas variáveis num modelo deve seguir critérios: incluir variáveis de controle irrelevantes afeta a qualidade da inferência. De acordo com Neumayer e Plümper (2017, p.123), a melhor forma de proceder não é tentar construir modelos que deem conta de toda variação de Y, mas construir modelos que tenham bom desempenho em testes de robustez, por exemplo rodando testes que investiguem variáveis omitidas que podem estar correlacionadas às variáveis de interesse.

---

2 Parte 2 - Pós-graduação

Para responder esta parte, deve-se tomar o banco de dados que está junto com este arquivo de enunciado para a prova (Base_prova_parte2.csv).

A variável dependente é votos e as demais colunas são todas variáveis explicativas.

Deve-se selecionar uma amostra aleatória de tamanho igual a 500 para esse exercício. Isto deve ser feito utilizando dois comandos no R: set.seed() e o sample_n().

Para o primeiro comando, deve-se novamente utilizar o número USP como o seed. Para um número USP como o do exemplo anterior, deve-se digitar no início do código: set.seed(12345678) e em seguida sample_n(500).

```
set.seed(13485660)

base_nova <-
  base %>%
  sample_n(500)
```

```
## sample_n: removed 9,500 rows (95%), 500 rows remaining
```

```
dim(base_nova)
```

```
## [1] 500   6
```

De posse deste novo banco de dados, deve-se buscar encontrar um modelo de bom ajuste para a previsão dos votos. O livro de código das variáveis está resumido a seguir:

- Votos: Votação total recebida por um candidato a deputado federal na eleição no ano t.
- financ: Valor (em milhares) declarado pelo candidato da arrecadação da campanha,
- idade: Idade do candidato (em anos) no dia 31/07 do ano t.
- Fpart: Proporção dos recursos de campanha do candidato recebida do partido.
- Votos_t1: Votação total recebida por um candidato a deputado federal na eleição no ano t-1.

Em sua resposta, você deve apresentar o modelo de MQO que melhor se adéqua aos dados selecionados. Se for necessário, considere a situação de apresentar mais de um modelo. Apresente também gráficos e estatísticas descritivas das variáveis que contribuam na sua decisão sobre qual o melhor modelo. Teste para a validade das hipóteses do modelo de MQO e mostre os resultados obtidos. Apresente ao final as primeiras 25 linhas do seu banco de dados também.

2.1 Estabelecendo hipóteses

Com a variável dependente sendo Votação total recebida para deputado federal, as primeiras hipóteses que podemos estabelecer dizem respeito a recursos financeiros: é esperado que um maior financiamento de campanha, bem como uma proporção maior de financiamento fornecida ao candidato em relação a outros de seu partido, influenciem num maior número de votos. Portanto podemos estabelecer como primeira hipótese:

> H1: as variáveis financ e Fpart terão coeficiente positivo e estatisticamente significativos. Isto é, candidatos com mais recursos de campanha recebem mais votos.

Em relação à idade do candidato, é esperado que candidatos mais velhos recebam mais votos, devido a motivos como um maior tempo de carreira. Portanto podemos estabelecer que:

> H2: a variável idade terá coeficiente positivo e estatisticamente significativo. Isto é, candidatos mais velhos recebem mais votos.

Por fim, podemos pressupor também que candidatos que já receberam um certo número de votos em eleição anterior tenderão a manter um número semelhante na eleição avaliada pelo modelo. Portanto, podemos estabelecer que:

> H3: a variável Votos_t1 terá coeficiente positivo e estatisticamente significativo. Isto é, candidatos que receberam mais votos em eleição anterior recebem mais votos nessa eleição.

Primeiramente, será realizada uma transformação em votos, votos no ano t-1 e financiamento, realizando divisão em ambas por mil. Em seguida, são construídos modelos bivariados e multivariados:

	Dependent variable:			
	votos_por_mil			
	(1)	(2)	(3)	(4)
arrecadação de campanha (por mil)	−0.22			
	(0.25)			
idade		3.50***		
		(0.35)		
votos recebidos em eleição t-1 (por mil)			1.19***	
			(0.06)	
Proporção dos recursos recebidos do partido				−47.39**
				(21.39)
Constant	213.81***	52.77***	176.71***	221.10***
	(3.48)	(16.16)	(2.66)	(4.92)
Observations	500	500	500	500
R^2	0.002	0.17	0.42	0.01
Adjusted R^2	−0.0005	0.16	0.42	0.01
Residual Std. Error (df = 498)	56.41	51.57	42.98	56.18
Note:				*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:		
	votos_por_mil		
	(1)	(2)	(3)
arrecadação de campanha (por mil)	−0.06	−0.10	−0.09
	(0.23)	(0.16)	(0.16)
idade	3.50***	3.68***	3.68***
	(0.35)	(0.24)	(0.24)
votos recebidos em eleição t-1 (por mil)		1.21***	1.20***
		(0.05)	(0.05)
Proporção dos recursos recebidos do partido			−18.14
			(13.61)
Constant	53.60***	9.99	13.60
	(16.51)	(11.54)	(11.84)
Observations	500	500	500
R^2	0.17	0.60	0.61
Adjusted R^2	0.16	0.60	0.60
Residual Std. Error	51.62 (df = 497)	35.59 (df = 496)	35.56 (df = 495)
Note:			*p<0.1; **p<0.05; ***p<0.01

2.2 Interpretando os modelos

A variável que se manteu mais estável foi a idade e votos recebidos na eleição t-1, mantendo coeficientes praticamente inalterados com a introdução de novas variáveis. A variável de financiamento não apresentou significância estatística em nenhum modelo, e a variável de porcentagem de recursos recebidos do partido perdeu significância esatística com a introdução de novas variáveis. A respeito do R^2, temos que o modelo com votos recebidos em eleição anterior é o que explica maior proporção da variação na variável dependente, explicao 42%. Em relação aos modelos multivariados, o modelo 02 possui 60% de poder de explicação da variável dependente.

É possível verificar se esses coeficientes realmente mudaram adicionando o intervalo de confiança na tabela de regressão:

	Dependent variable:		
	votos_por_mil		
	(1)	(2)	(3)
arrecadação de campanha (por mil)	−0.06	−0.10	−0.09
	(−0.50, 0.39)	(−0.41, 0.21)	(−0.40, 0.22)
idade	3.50***	3.68***	3.68***
	(2.80, 4.19)	(3.20, 4.15)	(3.20, 4.16)
votos recebidos em eleição t-1 (por mil)		1.21***	1.20***
		(1.11, 1.31)	(1.10, 1.31)
Proporção dos recursos recebidos do partido			−18.14
			(−44.82, 8.55)
Constant	53.60***	9.99	13.60
	(21.24, 85.97)	(−12.62, 32.60)	(−9.61, 36.81)
Observations	500	500	500
R^2	0.17	0.60	0.61
Adjusted R^2	0.16	0.60	0.60
Residual Std. Error	51.62 (df = 497)	35.59 (df = 496)	35.56 (df = 495)
Note:			*p<0.1; **p<0.05; ***p<0.01

Podemos ver que os valores do parâmetro idade estão contidos nos intervalos de um modelo para o outro, assim como votos recebidos em eleição t-1

Com a transformação na unidade de algumas variáveis, é necessário atentar-se para a interpretação. Dado que agora a variável dependente foi dividida por mil, no modelo 03 temos que a cada ano mais velho um candidato tende a receber 3.68×10^3 votos, ou 368000 votos. Já para a variável de votos na eleição t-1, a cada mil votos (dado que ambas as variáveis foram divididas por mil) recebidos em eleição anterior o candidato tende a receber 1.20×10^3 votos, ou 120000 votos.

Podemos realizar testes para verificar se os modelos são homocedásticos e se há problemas de multicolinearidade entre as variáveis:

2.3 Testes de heterocedasticidade e multicolinearidade

Utilizando o teste de Breusch-Pagan e o índice de VIF, podemos testar heterocedasticidade e multicolinearidade nos modelos:

Modelos Bivariados:

Table 1: P-valor de testes de Breusch-Pagan para os Modelos Bivariados

Modelo	P-valor
financ	0.9236557
idade	0.3402370
votos em t-1	0.3187596
fpart	0.3187596

Modelos Multivariados:

Table 2: P-valor de testes de Breusch-Pagan para os Modelos Bivariados

Modelo	P-valor
idade_financ	0.6185658
idade_financ_t1	0.0722291
completo	0.0751955

Teste de multicolinearidade:

Table 3: Teste de Variance inflation factor

Variáveis	Tolerância	VIF
financ_por_mil	0.9940489	1.005987
idade	0.9939799	1.006056
votos_t1_por_mil	0.9893136	1.010802
fpart	0.9896237	1.010485

O teste de VIF não indica multicolinearidade entre as variáveis. Utilizando o teste de Breusch-Pagan, não rejeitamos a hipótese nula do modelo em que a variância dos resíduos está distribuída de maneira igual. Ou seja, em todos os modelos é possível verificar homocedasticidade.

Podemos agora criar um modelo apenas com as variáveis que demonstraram significância estatística:

2.4 Modelo final:

Com o modelo final podemos rejeitar a H1, apontando que as variáveis de financiamento de campanha não foram estatisticamente significantes nos modelos. Podemos confirmar nossas hipóteses 02 e 03, apontando que as variáveis de idade e votos recebidos em eleição anterior t-1 tiveram significância estatística e coeficiente positivo, ou seja, candidatos mais velhos tendem a receber mais votos e candidatos que receberam mais votos em eleição anterior tendem a receber mais votos nessa eleição.

	Dependent variable:
	votos_por_mil
idade	3.69***
	(0.24)
votos recebidos em eleição t-1 (por mil)	1.21***
	(0.05)
Constant	8.57
	(11.30)
Observations	500
R^2	0.60
Adjusted R^2	0.60
Residual Std. Error	35.57 (df = 497)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.5 Referências

Neumayer, Eric and Thomas Plümper. 2017. Robustness Tests for Quantitative Research (Methodological Tools in the Social Sciences). Cambridge; New York: Cambridge University Press.

Kellstedt, Paul M., and Guy D. Whitten. 2015. Fundamentos da Pesquisa em Ciência Política (Lorena Barberia, Gilmar Masiero and Patrick Cunha Silva, Translators). São Paulo, Brazil: Editora Blucher.

2.6 Primeiras 25 linhas da base:

Table 4: Primeiras 25 linhas da base gerada com o comando set.seed utilizando NUSP 13485660

votos	financ	idade	n_part	fpart	votos_t1	votos_por_mil	financ_por_mil	votos_t1_por_mil
155161.3	16562.533	40	0	0.160	2613.365	155.161	16.563	2.613
257347.1	2762.696	44	3	0.499	51688.781	257.347	2.763	51.689
328636.0	4622.792	40	1	0.146	158642.000	328.636	4.623	158.642
203057.0	1396.822	41	2	0.210	8217.878	203.057	1.397	8.218
202156.1	7649.180	42	0	0.173	12825.900	202.156	7.649	12.826
205424.9	12206.756	52	3	0.251	6418.233	205.425	12.207	6.418
234903.0	10095.696	55	3	0.197	42413.844	234.903	10.096	42.414
153157.0	28723.135	40	3	0.179	31315.555	153.157	28.723	31.316
205822.4	920.206	45	1	0.118	12252.076	205.822	0.920	12.252
179141.1	6948.481	54	3	0.172	11101.063	179.141	6.948	11.101
303444.7	12151.393	47	3	0.124	80744.391	303.445	12.151	80.744
205936.5	9723.636	41	2	0.239	20492.201	205.936	9.724	20.492
154774.5	3517.503	34	5	0.144	43645.180	154.774	3.518	43.645
174135.7	13903.864	35	2	0.346	33721.492	174.136	13.904	33.721
134693.8	658.522	45	0	0.204	11485.117	134.694	0.659	11.485
305120.5	3035.540	59	1	0.092	47856.781	305.120	3.036	47.857
267191.8	3490.525	59	4	0.030	28580.969	267.192	3.491	28.581
334055.4	17233.699	42	3	0.201	96128.109	334.055	17.234	96.128
335747.2	2987.853	51	3	0.527	68426.992	335.747	2.988	68.427
255221.8	10303.804	47	1	0.105	35506.801	255.222	10.304	35.507
220827.8	1025.080	47	2	0.129	21363.342	220.828	1.025	21.363
174830.8	3080.033	44	0	0.183	6083.706	174.831	3.080	6.084
140612.3	2723.677	43	0	0.438	10885.847	140.612	2.724	10.886
219342.6	19539.031	48	2	0.102	20250.844	219.343	19.539	20.251
219732.7	6521.145	48	4	0.097	16012.951	219.733	6.521	16.013