**Universidade de São Paulo**

**Faculdade de Filosofia, Letras e Ciências Humanas**

**Departamento de Ciência Política**

Prova Parcial - Amanda Freitas – NUSP 13485660

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

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

13485660

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.

1. 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% | | | | |

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 \* X1) + ε |

Neste modelo, quando X1 é igual a 0, Y é igual a 2,10. A cada aumento de 1 unidade na variável independente X1, a variável dependente Y aumenta em 1 unidade.

1. 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 X1 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 X1 não está associada à variável dependente Y. É possível realizar o teste de hipóteses por meio de Intervalo de Confiança, à 99% de significância estatística, tendo como hipótese nula que o coeficiente de X1 passe por 0, isto é, não está associado à Y:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Parâmetro de X1 | Erro-padrão | IC (99%) |
| Modelo 1 | 1,30 | 0,02 | 1,24 ; 1,36 |

Como 0 não está contido no Intervalo de Confiança, podemos rejeitar a hipótese nula.

1. 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, 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.

1. 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 | Erro-padrão | IC (99,7%) | IC (95%) | IC (68%) |
| Modelo 1 | 2,10 | 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.

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

|  |
| --- |
| Y = 1,54 + (1,42 \* X1) + (-4,12 \* X2) + ε |

1. 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 X1 mudou com a introdução de X2 a partir do intervalo de confiança. No modelo 1, temos um resultado de (beta) para X1 igual a 1,30, com 0,02 de erro-padrão. Já no modelo 2, o parâmetro de X1 mudou para 1,42, com um erro-padrão de 0,01. Adicionando 2 erros-padrão e subtraindo 2 erros-padrão, temos o seguinte intervalo de confiança:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Parâmetro de X1 | Erro-padrão | IC |
| Modelo 1 | 1,30 | 0,02 | 1,26 ; 1,34 |
| Modelo 2 | 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.

1. O que significa o termo R2? Interprete o valor encontrado para o segundo modelo;

O termo R2 indica a fração de variação da variável dependente Y que é explicado pela variável independente 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 R2, 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 R2 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.

1. 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, o (beta) de X2 era igual a -4,12, com erro padrão igual a 1,90 e significância estatística ao nível de 95%. Já com a introdução da variável X3, o erro padrão aumentou para 6,70; além disso, o coeficiente de X2 foi para 6,60 e mudou de sinal. Isso indica que pode haver problema de multicolinearidade com as variáveis independentes, o que viola as hipóteses necessárias para o modelo de regressão linear. Um R2 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.

1. 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.

1. 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.

**2ª parte – 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)*.

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:

|  |  |
| --- | --- |
| Variável | Significado |
| 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.