Universidade Federal do Pará
Programa de Pós-graduação em Ciência Política
Disciplina: Tópicos Especiais Em Ciência Política:
Pesquisa Quantitativa em Ciência Política
Créditos: 4
Carga horária: 60h

AULA 11
MODELOS LINEARES

PROFA. NAIARA ALCANTARA E LUCAS OKADO

2023

RESUMO

INTRODUÇÃO A REGRESSÃO

1. Gráfico de dispersão
2. Regressão linear simples
3. Análise de resíduos

REGRESSÃO LINEAR MÚLTIPLA

1. Teste
2. Condicionantes
3. Apresentação gráfica

REGRESSÃO LINEAR

- O modelo de regressão linear simples explica uma variável (y) com base em modificações em outra variável (x).
- Ou seja, é usado para avaliar a relação entre duas variáveis.

*Estamos interessados em explicar y, através da observação de x.
Ou seja, queremos estudar alterações em Y com base nas variações de x.

1- Quero explicar o rendimento salarial (Y). O que pode explicar ?

X=Networking
2- Quero explicar o conservadorismo religioso (Y). O que pode explicar?
X= Denominação religioso/ X=Frequência religiosa/ X=Literalismo bíblico

REGRESSÃO LINEAR SIMPLES

- A mensuração do teste ocorre através de uma estimativa, que assim como outros testes, tem uma variação que vai de -1 a +1, passando por 0.
- Quanto mais próximo de 0,00 menor é a evidência de correlação. Valores próximo de -1 indicam correlação negativa e próximos de 1 revelam correlação positiva.
- Na prática esse é um modelo estatístico que busca explorar a relação entre uma variável dependente e uma variável independente*
- **Em modelos lineares a variável dependente sempre terá que ser numérica**

y	x
Variável Dependente	Variável Independente
Variável Explicada	Variável Explicativa
Variável de Resposta	Variável de Controle
Variável Prevista	Variável Previsora

Existem algum pressupostos que teremos que responder, para saber se o modelo de regressão linear é adequado para os dados.

- A melhor forma de entender se há relação entre as variáveis é através de um gráfico de dispersão.

REGRESSÃO LINEAR SIMPLES – NA PRÁTICA

As análises serão realizadas utilizando o **banco** de candidatos eleitos e não eleitos para o senado no ano de 2018

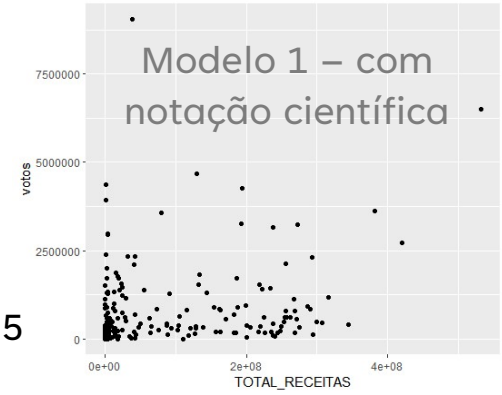
Uma boa forma de avaliar se existe relação entre as variáveis é através de um gráfico de dispersão.

Eixo y: Votos

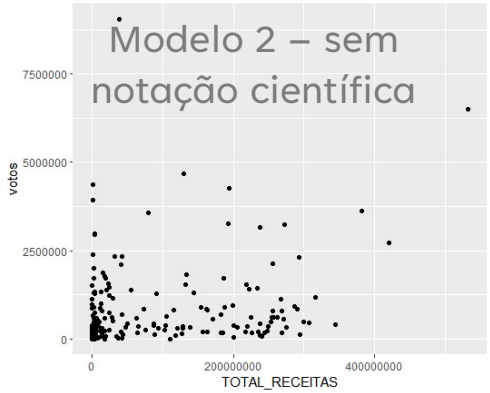
0

Eixo x: Receita

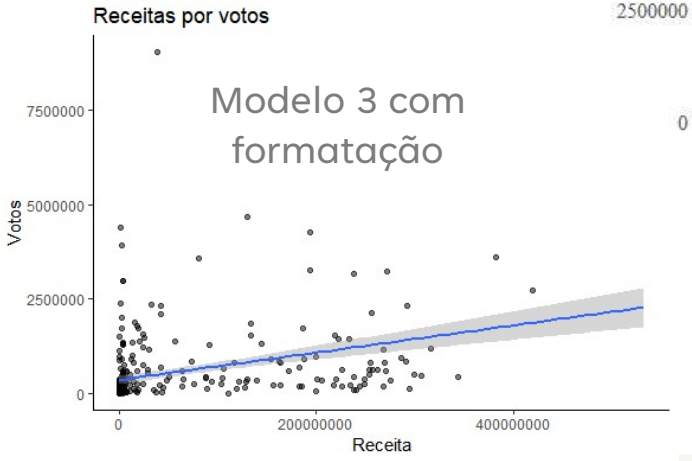
Modelo 1 – com notação científica



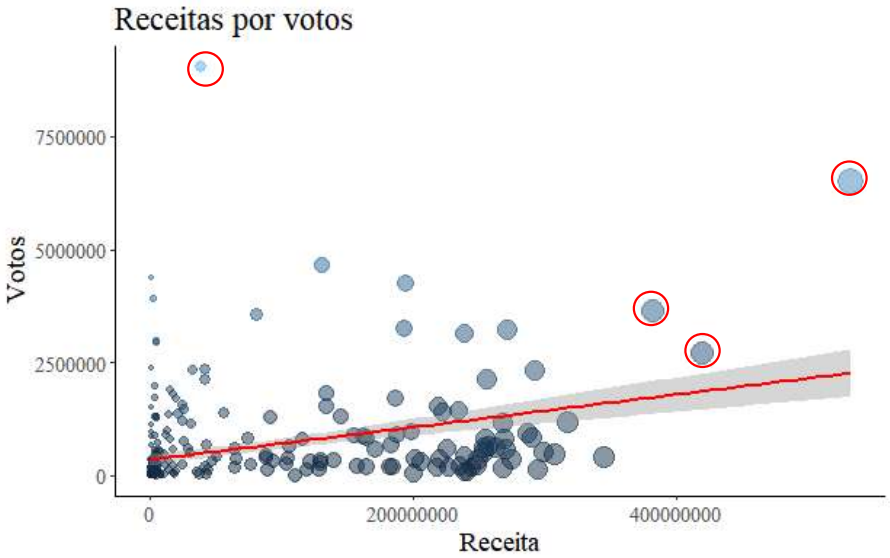
Modelo 2 – sem notação científica



Modelo 3 com formatação



Modelo 4 final



REGRESSÃO LINEAR SIMPLES – NA PRÁTICA

```
> summary(Model1.1)

Call:
lm(formula = log.votos ~ log.receitas, data = sen2018)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3221 -0.7835  0.0096  0.7186  4.3688

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   5.52857    0.34657   15.95 <0.0000000000000002 ***
log.receitas  0.53285    0.02792   19.09 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.175 on 300 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.5484,    Adjusted R-squared:  0.5469
F-statistic: 364.3 on 1 and 300 DF,  p-value: < 0.00000000000000022
```

Saída do modelo usando a função summary

O residual standard error é uma medida da dispersão dos resíduos em torno da linha de regressão. Neste caso, é 1.366.

Resíduos: servem para indicar pontos influentes (distantes) e a tendência geral de acerto do modelo

Coeficientes: O intercepto de **7,43** é estatisticamente significativo (p <0.001) e indica que um candidato com log de gasto 0 teria o log de votação de 7.43. O coeficiente de log.vgasto é estatisticamente significativo e indica que a cada ponto adicionam nesse medida ocorre a elevação de 0,30 na votação

Legenda dos níveis de significância das estrelinhas

Estatísticas de ajuste do modelo: o R-quadrado indica a proporção de variação na variável dependente que pode ser explicada pela(s) variável(s) independente(s). O R-quadrado tem como valor máximo 1, portanto **cerca de 38% dos votos são explicados pela variação das receitas.**

REGRESSÃO LINEAR SIMPLES – NA PRÁTICA

Saída do modelo usando a função tab_model

```
tab_model(Model1.1, show.ci = F, auto.label = T, show.se =  T,
          collapse.se = T, wrap.labels = 60, p.style = "stars")
```

	log.votos
Predictors	*Estimates*
(Intercept)	5.53 *** (0.35)
log receitas	0.53 *** (0.03)
Observations	302
R^2 / R^2 adjusted	0.548 / 0.547

*$p<0.05$ **$p<0.01$ ***$p<0.001$

Equação da reta
Y = a + bX + e
Y = 7,53 + 0,53*X + e

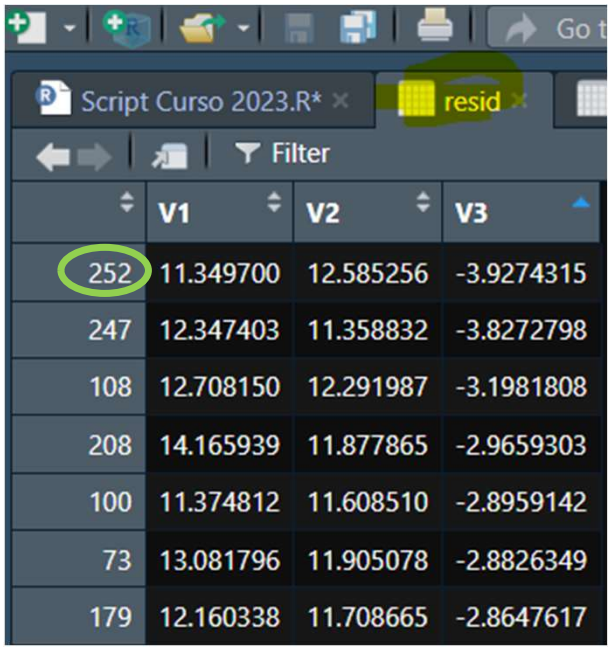
ITENS QUE COMPÕE O TAB_MODEL

- show.se = (mostra o desvio padrão) se não colocarmos ele não aparece
- collapse.se = (esse item vai juntar "colapsar" o desvio padrão Std e a estimativa, isto é, o valor do desvio irá aparecer logo em baixo do valor da estimativa)
- auto.label = única função desse item é separar a variável da categoria que ela está testando. Por exemplo, a variável é faixa etária e está testando para idoso, na saída irá aparece Faixa etária [idoso]).
- Show.ci = (Mostra o intervalo de confiança (Ci)) devemos pedir sempre que não apareça, usando o F
- Quando pedimos os coeficientes padronizados de beta, podemos colocar no tab_model essa função (show.std = T), para verificar o desvio padrão do beta
- Por fim, os estilos podem ser os seguintes:"numeric", "stars", "numeric_stars", "scientific", "scientific_stars"

DIAGNÓSTICOS DE RESÍDUOS

- A análise de regressão pode ser influenciada por pontos influentes (outliers) que podem fazer com que os resíduos se elevem bastante;
- Além disso, algumas exigências quanto à distribuição residual precisam ser confirmadas
- Isso pode ser feito construindo uma planilha com os valores preditos e os erros para cada caso a partir do nosso modelo
- Comando cbind que combina linhas e colunas em uma base de dados

```
#Análise de resíduos
resid ← (cbind(sen2018$log.votos, predict(Model1.1),
                residuals(Model1.1)))
view(resid)
```



	V1	V2	V3
252	11.349700	12.585256	-3.9274315
247	12.347403	11.358832	-3.8272798
108	12.708150	12.291987	-3.1981808
208	14.165939	11.877865	-2.9659303
100	11.374812	11.608510	-2.8959142
73	13.081796	11.905078	-2.8826349
179	12.160338	11.708665	-2.8647617

	SQ_CANDIDATO	SIGLA_UF	NOME_CANDIDATO
252	230000601676	RR	ROMERO JUC FILHO

votos	TOTAL_RECEITAS	DESC_SIT_TOT_TURNO
84940	2250000	N O ELEITO

R$2.250.000

SLIDE OCULTO DOS CANDIDATOS QUE GANHARAM

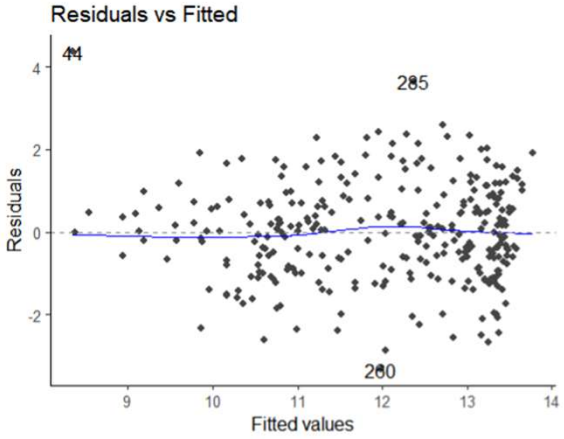
SQ_CANDIDATO	SIGLA_UF	NOME_CANDIDATO	NUMERO_PARTIDO	SIGLA_PARTIDO	DESCRICAO_COR_RACA	DESCRICAO_OCUPACAO	DESC_SIT_TOT_TURNO	SITUACAO_REELEICAO	DES_SITUACAO_CANDIDATURA	votos	TOTAL_RECEITAS
2,3E+11	RR	FRANCISCO DE ASSIS RODRIGUES	25	DEM	PARDA	ENGENHEIRO	ELEITO	N	APTO	111466	891508,3
2,3E+11	RR	ANTÔNIO MECIAS PEREIRA DE JESUS	10	PRB	PRETA	ADMINISTRADOR	ELEITO	N	APTO	85366	2380209

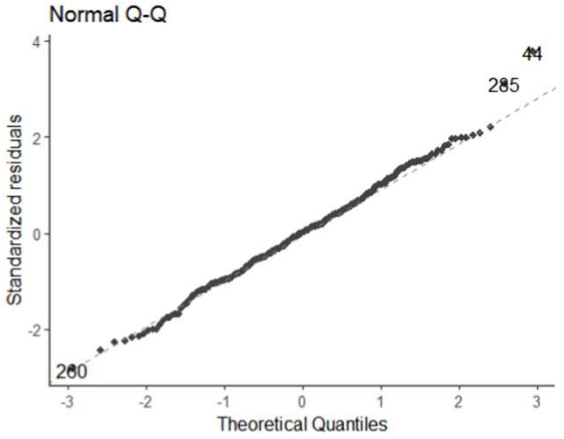
ANÁLISE DOS RESÍDUOS

Outra forma de olhar para os resíduos é através de gráficos.

Assumimos que os dados são aproximadamente normalmente distribuídos

1º resíduos pelos valores ajustados, linearidade, aproximadamente horizontal útil para testarmos a independência entre valores preditos e resíduos

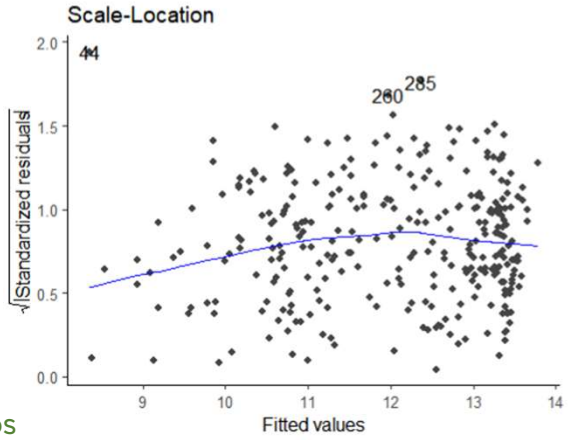


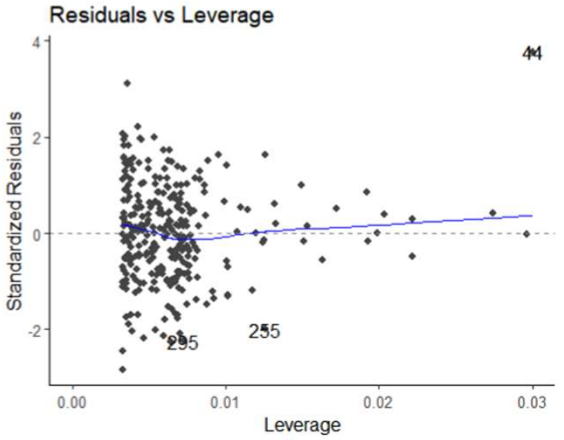


2º Distribuição normal ou Normal q –q (Quantile-Quantile)

Os pontos devem estar deve estar em cima da linha. Nos ajuda a verificar essa exigência ao exibir no eixo horizontal a distribuição esperada em uma distribuição normal e no vertical os resíduos padronizados

3º O Scale-localtion é uma variação do gráfico "Residuals vs Fitted" e é útil para verificar a homogeneidade da variância dos resíduos, que é uma das suposições-chave da análise de regressão linear .





4º Residuals vs leverage (Resíduos vs Alavancagem) é utilizado para identificar observações que podem ter um impacto desproporcional na estimativa dos coeficientes do modelo de regressão .

os pontos estão espalhados aleatoriamente em torno da linha de referência, sugere que a suposição de homogeneidade da variância está sendo atendida.

Não existem muitos pontos de outliers que estão tanto distantes no eixo x (alta alavancagem) quanto distantes no eixo y (alto resíduo ou resíduo studentizado), são pouquíssimos casos indicados em cada uma das três projeções gráficas, sendo assim não faremos análise de cada caso.

Análise multivariada: É muito semelhante ao modelo simples, porém com a inserção de mais variáveis explicativas.

Geralmente uma questão social é explicada apenas por um fator?

	log.votos	
Predictors	*Estimates*	*p*
(Intercept)	6.38 *** (1.50)	**<0.001**
log receitas	0.52 *** (0.03)	**<0.001**
Idade	-0.01 (0.01)	0.402
Sexo [MASCULINO]	-0.10 (0.18)	0.574
Instrução [Média]	-0.58 (0.86)	0.505
Instrução [Superior]	-0.57 (0.84)	0.495
Raca [BRANCA]	0.35 (1.17)	0.766
Raca [INDÍGENA]	-0.97 (1.44)	0.501
Raca [PARDA]	0.00 (1.18)	0.997
Raca [PRETA]	-0.15 (1.19)	0.900
Observations	302	
R^2 / R^2 adjusted	0.564 / 0.551	

*$p<0.05$ **$p<0.01$ ***$p<0.001$

Interpretação

1-As variáveis que acreditávamos ser importante foram de fato?

2- A capacidade preditiva do teste mudou?

CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

1º Colinearidade

Definição: a perfeita colinearidade ocorre quando uma das variáveis independentes possui uma relação linear perfeita com outra (s) variável(s) independentes.

Problemas:

- Se as duas variáveis se relacionam intimamente provavelmente são redundantes, ou seja, medem a mesma coisa, sendo impossível distinguir o efeito de uma e outra.

Diagnóstico de Colinearidade

1º Variance Inflation Factors

2º Condition Index

CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

1º Variance Inflation Factors

- Mede a inflação na estimativa de um parâmetro derivada da
colinearidade existente entre os preditores
- VIF de 1 indica que não há correlação entre um determinado preditor e os demais incluídos no modelo, logo, não ocorre a inflação do seu coeficiente estimado
- Valores próximos de 4 inspiram cuidado e próximos de 10
revelam grave problema

Pacote olsrr
*script

	Variables	Tolerance	VIF
1	log.receitas	0.85347947	1.171674
2	Idade	0.87446042	1.143562
3	SexoMASCULINO	0.95858347	1.043206
4	InstruçãoMédia	0.06797869	14.710492
5	InstruçãoSuperior	0.06854282	14.589419
6	RacaBRANCA	0.01498990	66.711593
7	RacaINDÍGENA	0.33168427	3.014915
8	RacaPARDA	0.01971043	50.734548
9	RacaPRETA	0.03379654	29.588826

2º Condition Index

-Uma alternativa ao VIF é o Condition Index que se baseia no cálculo dos eigenvalues para cada variável, que basicamente indicam o quanto da sua variabilidade está relacionada a outra variável do modelo

- Eigenvalues próximos de “0” indicam forte colinearidade e o condition index é basicamente a raíz quadrada do maior eigenvalue para cada variável

- Condition index com valor maior do que 30 alertam para problema grave

```
> ols_eigen_cindex(Model2)
   Eigenvalue Condition Index
1  5.782131890         1.000000
2  1.134681419         2.257392
3  1.002991579         2.401018
4  1.001966546         2.402245
5  0.866100876         2.583805
6  0.150537796         6.197567
7  0.029128592        14.089132
8  0.025985467        14.916905
9  0.005128573        33.577298
10 0.001347264        65.511519
```

CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

Diante da constatação de que existe multicolinearidade entre os preditores algumas estratégias podem ser adotadas.

- Aumentar o tamanho da amostra nos casos em que existem poucos casos e muitas variáveis no modelo.
- Combinar preditores em alguma espécie de indicador ou índice.
- Combinar as raças em uma única ou em duas como branca e parda
- Excluir variáveis redundantes.

CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

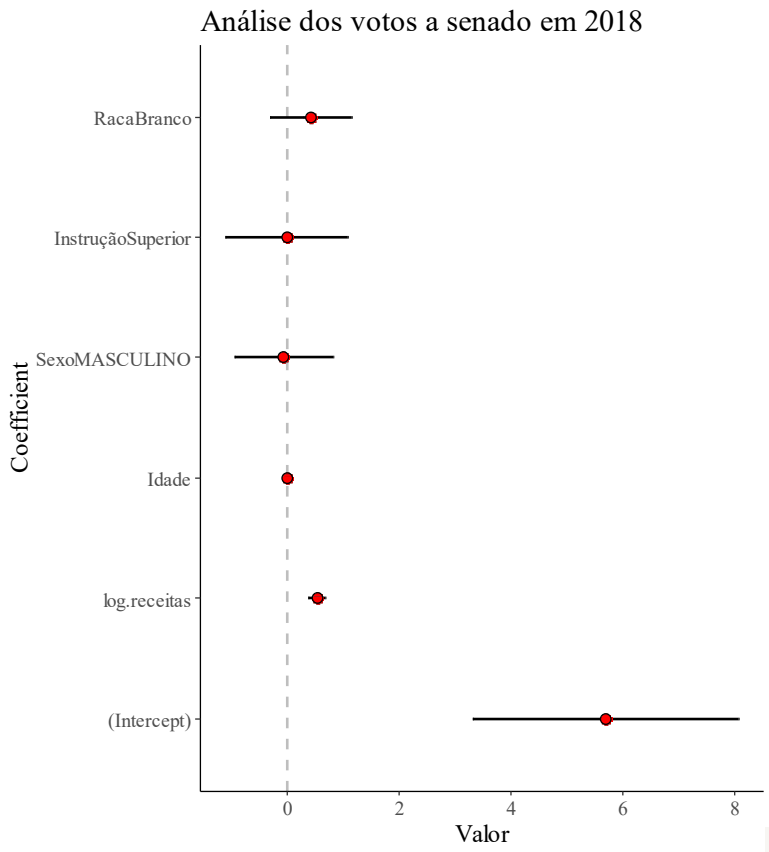
	log.votos	
Predictors	*Estimates*	*p*
(Intercept)	5.69 *** (0.48)	**<0.001**
log receitas	0.52 *** (0.03)	**<0.001**
Idade	-0.00 (0.01)	0.446
Sexo [MASCULINO]	-0.07 (0.18)	0.679
Instrução [Superior]	-0.01 (0.22)	0.955
Raca [Branco]	0.42 ** (0.15)	**0.005**
Observations	302	
R^2 / R^2 adjusted	0.561 / 0.554	

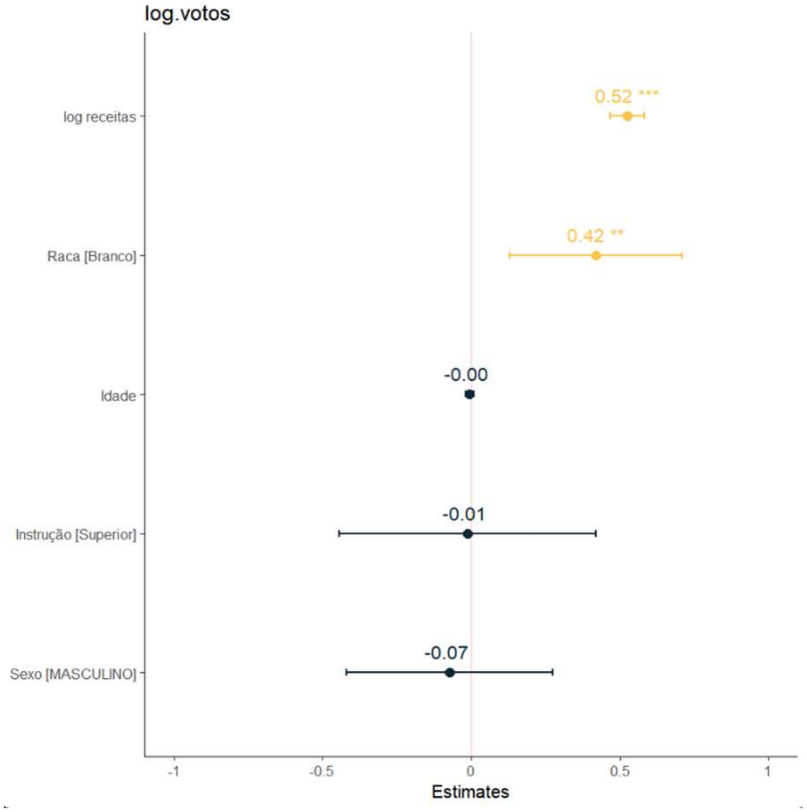
** p<0.05 ** p<0.01 *** p<0.001*

```
> ols_vif_tol(Model2)
          Variables Tolerance      VIF
1     log.receitas 0.8823612 1.133323
2            Idade 0.8819884 1.133802
3    SexoMASCULINO 0.9722859 1.028504
4 InstruçãoSuperior 0.9880685 1.012076
5       RacaBranco 0.9415318 1.062099
> #Condition Index####
> ols_eigen_cindex(Model2)
  Eigenvalue Condition Index
1 5.43474208        1.000000
2 0.29239198        4.311285
3 0.14895897        6.040267
4 0.08256944        8.112969
5 0.02662812       14.286281
6 0.01470939       19.221713
```

APRESENTAÇÃO GRÁFICA DO TESTE DE REGRESSÃO

O Coefplot é um tipo de gráfico que pode ser utilizado para apresentar graficamente os resultados de um teste de regressão





OBRIGADA ! =)