LABORATÓRIO 2: Regressão Linear Simples

Fernando Bispo, Jeff Caponero

Sumário

Introdução

O presente relatório está subdividido em duas partes, tendo em vista terem sido disponibilizados dois arquivos para análise, este relatório vem trazendo as análises conforme os pré-requisitos solicitados para cada conjunto de dados. Tendo como principal objetivo a introdução das técnicas de Regressão Linear Simples e a pratica da elaboração de relatórios analíticos fundamentadas na Análise Exploratória de Dados.

Primeira parte

Apresentação

Nesta primeira parte a análise se dará sobre os dados contendo medidas morfológicas de 104 gambás das montanhas, que foram capturados através de armadilhas em sete localizações na Inglaterra. As variáveis contidas nesse arquivo são:

- Sexo (**sex**);
- Largura do crânio (**skullw**);
- Comprimento total (**totlngth**).

Objetivo

O objetivo dessa análise visa responder aos seguintes tópicos:

(a) Descrição do comportamento de cada uma das variáveis, a partir das medidas morfológicas segundo o sexo.

(b) Representação gráfica da distribuição do sexo.

(c) Apresentação de um histograma para as variáveis morfológicas.

(d) Discução da relação entre as variáveis morfológicas.

(e) Tratamento dos dados.

(f) Avaliação do ajuste de um modelo linear de regressão.

(g) Caso o ajuste seja adequado, apresentar a reta ajustada pelo modelo.

Análise dos dados

Composto por três características (variáveis) morfológicas dos gambás, em que duas destas são classificadas como **aritméticas continuas**, sendo estas a largura do crânio (**skullw**) e o comprimento total (**totlngth**) e uma variável classificada como **categórica ordinal**, sendo esta sexo(**sex**).

A seguir são apresentadas as tabelas com as principais medidas resumo por sexo dos gambás.

Table 1: Medidas Resumo para o sexo feminino.

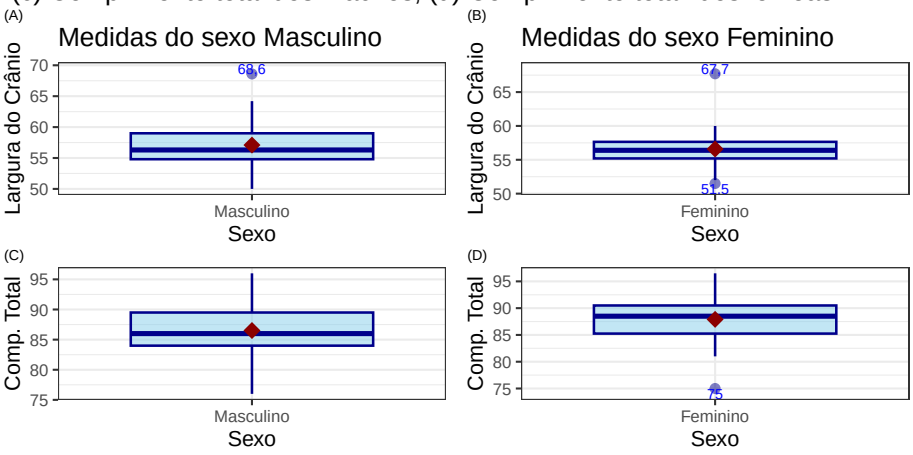
	Min	Q1	Med	Média	Q3	Max	D. Pad	CV
Comp. Total	75,0	85	88,5	87,91	90,5	96,5	4,18	0,05
Larg. Crânio	51,5	55	56,4	56,59	57,7	67,7	2,57	0,05

Table 2: Medidas Resumo para o sexo masculino

	Min	Q1	Med	Média	Q3	Max	D. Pad	CV
Comp. Total	76	84,0	86,0	86,51	89,5	96,0	4,34	0,05
Larg. Crânio	50	54,8	56,3	57,09	59,0	68,6	3,45	0,06

Nota-se que as medidas de resumo não apresentam diferenças significativas entre os dois sexos, avaliando o Coeficiente de Variação de Pearson (CV), sendo esta a medida que avalia o grau de variabilidade dos dados em relação a média, constata-se que sua classificação é baixa, possuindo o valor de 5%, menor que o limite considerado baixo (15%), para todas as características independente do sexo. A fim de se possibilitar uma análise visual, a Figura 1 traz os gráficos de caixa (*BoxPlot*) com as características morfológicas dos gambas separadas por sexo.

Figura 1: Comparação das medidas morfológicas por sexo.
(a) Largura do crânio nos machos; (b) Largura do crânio nas fêmeas;
(c) Comprimento total dos machos; (d) Comprimento total dos fêmeas



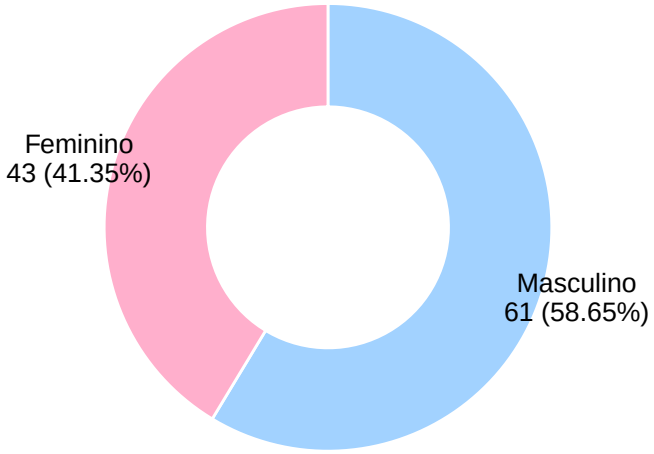
Nota:
Ponto vermelho ao centro do gráfico indicando a média dos dados.

A Figura 1 confirma a percepção advinda das medidas resumo, que não há, em princípio, diferenças significativas entre os sexos, apesar de haver alguns pontos atípicos (*outliers*). É interessante notar a baixa variabilidade entre a largura do crânio para o sexo feminino e seu comportamente aparentemente simétrico, bem como o fato de que apenas o comprimento dos individuos do sexo masculino não apresentou pontos atípicos.

Tendo em vista a média estar representada pelo ponto vermelho na região central do gráfico, possibilitando uma percepção visual acerca da assimetria dos dados ao comparar o posicionamentos da média e da mediana em relação a região central da caixa (região retangular azul da figura), que representa 50% dos dados. Cabe então uma avaliação mais apurada dos dados.

A Figura 2 retrata a distribuição do sexo, possibilitando identificar o quantitativo de informações por sexo.

Figura 2: Distribuição da variável sexo



Correspondendo ao item c dos objetivos da análise, segue a Figura 3 com os histogramas das características aritméticas dos gambás sob análise, sem subdivisão por sexo. A fim de identificar com maior clareza a distribuição dos dados das variáveis constantes nos histogramas seguem gráficos de densidade juntamente com as médias e medianas sobrepostas na Figura 4.

Figura 3: Histograma das medidas morfológicas.
(a) Largura do crânio; (b) Comprimento total.

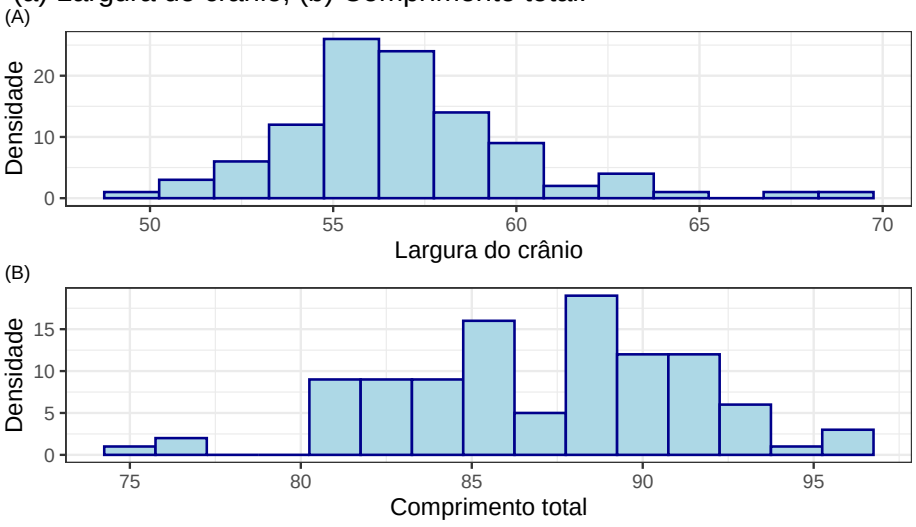
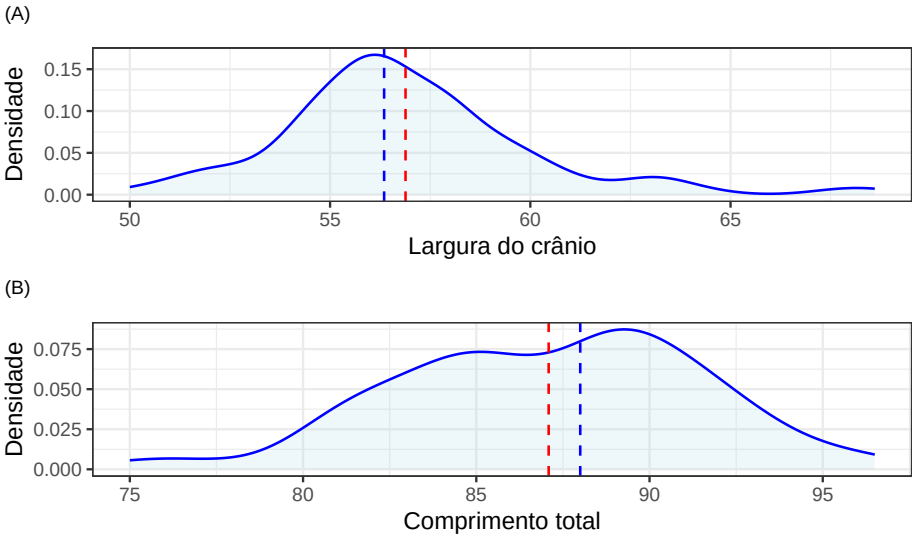


Figura 4: Densidade das medidas morfológicas dos gambás
(a) Largura do crânio; (b) Comprimento total.



Notas: Linha vertical tracejada azul representando a mediana dos dados.
Linha vertical tracejada vermelha representando a média dos dados.

Analisando as Figuras 3 e 4 é possível identificar uma leve assimetria positiva para os dados referentes a largura do crânio dos gambás, para os dados referentes ao comprimento total se identifica um comportamento bimodal e uma leve assimetria negativa. A fim de dirimir quaisquer dúvidas sobre a assimetria e curtose dos dados em análise, seguem as tabelas com os valores calculados.

Baseado nos valores que constam na Tabela 3, referente ao coeficiente de assimetria, é possível concluir que a distribuição dos dados referentes ao **comprimento total** apresenta um coeficiente de assimetria positivo, tendo em vista seu valor maior que 0, indicando que a maioria dos valores são menores que a média, caracterizada pela linha tracejada vertical vermelha presente nos gráficos de densidade, já os dados referentes a **largura do crânio** são considerados de assimetria negativa, ou seja, a maioria dos valores dos dados são maiores que a média, em virtude do seu valor menor que zero.

Table 3: Coeficientes de Assimetria

Variável	Coeficiente
Largura do Crânio	0,99
Comprimento Total	-0,28

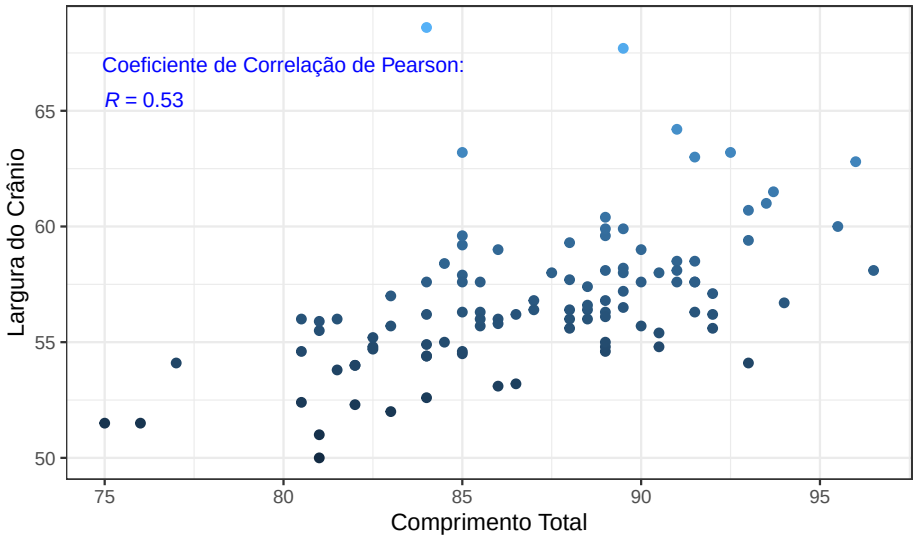
Table 4: Coeficientes de Curtose

Variável	Coeficiente
Largura do Crânio	5,30
Comprimento Total	2,84

Referente ao coeficiente de curtose (Tabela 4) é possível concluir que a distribuição dos dados referentes ao **comprimento total** possui valor menor que 3, caracterizando um comportamento Platocúrtico, ou seja, mais achatado, já a distribuição dos dados referentes a **largura do crânio** apresenta um coeficiente maior que 3, caracterizando um comportamento Leptocúrtico, demonstrando possuir um pico mais acentuado no gráfico, característica identificada nos gráficos de densidade.

Com o intuito de comparar a relação entre as variáveis largura do crânio (**skullw**) e comprimento total (**totlngth**), de forma visual, foi construída a Figura 5.

Figura 5: Relação entre Comprimento Total e Largura do Crânio



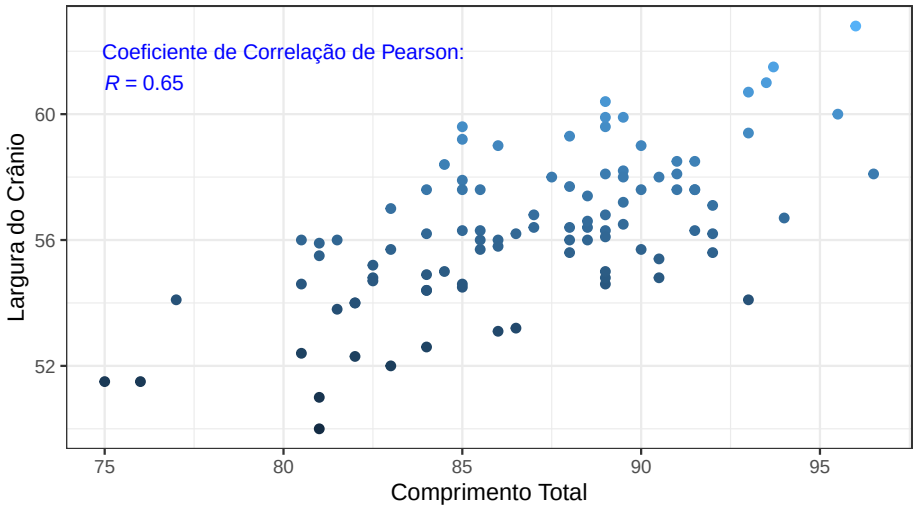
É possível identificar uma relação linear positiva entre as características em analise, ao se calcular o coeficiente de correlação de Pearson $\hat{\rho}$ estimado obteve-se o valor de 0.5264 caracterizando uma relação **regular**.

Para testar a significância da correlação linear ao nível de significância de 5% foi realizado o teste de hipótese para correlação linear, em que a hipótese nula foi rejeitada, podendo assumir com 95% de confiança que **há correlação linear entre as variáveis estudadas**.

Através da Figura 5 é possível identificar com mais clareza pontos atípicos na

relação entre as variáveis morfológicas, a fim de identificar o quão influente são estes pontos no comportamento geral dos dados, a Figura 6 foi construída sem estes pontos.

Figura 6: Relação entre Comprimento Total e Largura do Crânio com remoção dos pontos atípicos



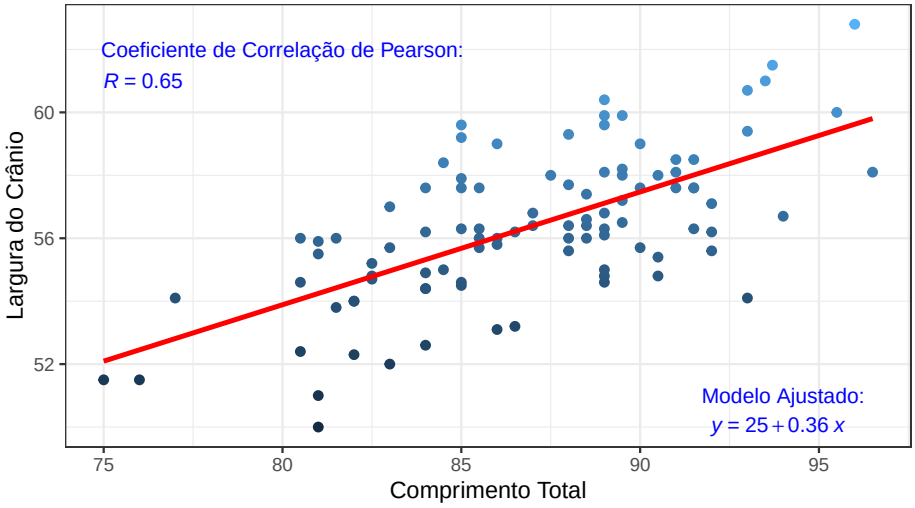
Analisando o comportamento da relação sem os pontos atípicos não se nota diferença significativa acerca do comportamento da relação, permanecendo uma possível relação positiva. Ao se calcular o novo coeficiente de correlação de Pearson $\hat{\rho}$ estimado obteve-se o valor de 0.6542, ligeiramente maior que o anteriormente calculado com todos os dados inseridos no cálculo, caracterizando ainda assim uma relação **regular**.

Em seguida foi feito o teste hipótese que avalia a significância da correlação linear ao nível de 5%, em que a hipótese nula foi rejeitada, podendo assumir com 95% de confiança que **permanece a correlação linear entre as variáveis estudadas**.

Tendo em vista a relação linear entre os dados em análise, é adequado o estudo que avaliará se a Largura do crânio pode ser explicada pelo Comprimento total do gambá, o que se vê na figura a seguir.

Verifica-se que é possível ajustar os dados a uma regressão linear com função $f(x) = 25 + 0,36x$ e com coeficiente de correlação de Pearson de 0,65.

Figura 7: Modelo de Regressão Ajustado entre
o Comprimento Total e Largura do Crânio



Segunda parte

Apresentação

Com base nos dados sobre a eleição presidencial de 2000 nos Estados Unidos, referentes ao número de votos de cada um dos candidatos por condado no estado da Flórida. Deseja-se investigar a relação entre o número de votos que Bush recebeu em relação ao número de votos recebidos por Buchanan, bem como, trazer um pouco de luz sobre o debate referente aos votos recebidos por Buchanan que poderiam ter sido de Al Gore, se o primeiro não estivesse no pleito. Como Bush e Gore foram os candidatos principais daquela eleição, é de interesse avaliar a relação entre os votos recebidos por Bush e Buchanan na Flórida, que é um Estado importante na corrida presidencial dos EUA. Para isto, ajuste um modelo de regressão linear no qual o número de votos de Bush é usado para predizer o número de votos de Buchanan.
Os dados estão disponíveis no arquivo "florida.csv".

Objetivos

O objetivo dessa análise responderá as seguintes questões:

(a) Discutir a relação entre os votos recebidos por Bush e por Buchanan através do uso de diagramas de dispersão.

(b) Avaliar a relevância do argumento que os votos de Buchanan poderiam ser direcionados a Al Gore, caso Buchanan não tivesse participado do pleito.

(c) Tratar dados atipicos.

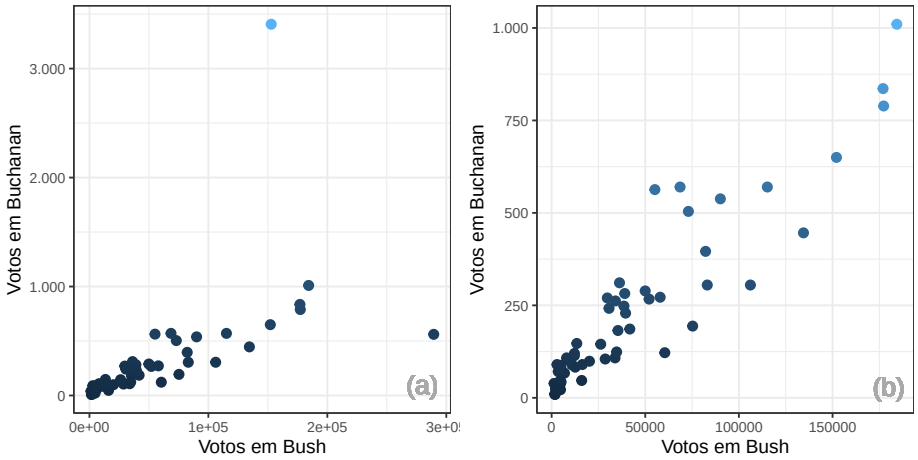
(d) Obter uma reta ajustada aos dados tratados e predizer a votação de Buchanan caso Bush receba 152.846 votos em determinado condado.

(e) Criar um programa baseado em estimativas de mínimos quadrados, prever a votação de Buchanan, sob as mesmas condições e compará-las.

Análise dos dados

O gráfico de dispersão sugere certa correlação positiva entre os votos de Bush e Buchanan, muito embora se observe que no condado de Palm Beach a votação de Buchanan (3.407 votos) represente um dado inesperado, bem como, em menor medida, a votação de Bush em Dade (289.456 votos). Retirados estes valores discrepantes, esta provável correlação parece ainda mais certa.

Figura 8: Relação entre votos recebidos por Bush e por Bachanan nos condados da Flórida. (a) Dados completos, (b) Sem dados discrepantes



Fonte: 2000 OFFICIAL PRESIDENTIAL GENERAL ELECTION RESULTS – USA

Embora a Figura 8 sugira certa correlação positiva entre os votos de Bush e Buchanan, a análise destes mesmos votos em relação a votação recebida por Al Gore e os votos totais dos condados parece retratar outra realidade (ver Figura 9), qual seja, que a correlação positiva é mais propriamente deviada a um fator externo, e não avaliado inicialmente, que corresponde ao aumento de eleitores nos condados. Desta forma, não parece ser plausível afirmar que a votação dada ao candidato Buchanan seria direcionada a qualquer dos candidatos caso este não participasse do pleito.

Uma maior compreenção da dispersão dos dados das votações dos candidatos, optou-se por analizar a variabilidade dos dados de votação dos candidatos. Na Tabela 5 é possível verificar uma maior variabilidade dos votos do candidato Bush em todas as análise realizadas, o que se deve principalmente a quantidade muito superior de votos recebidos.

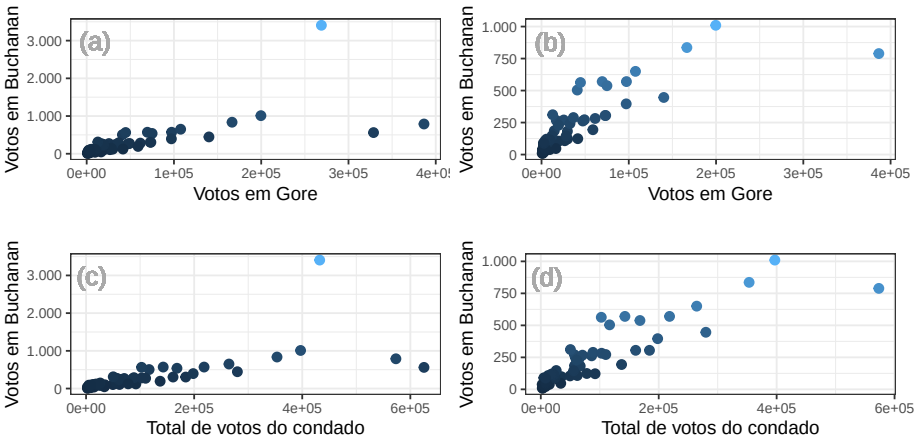
Table 5: Medidas de Variabilidade dos votos dos candidatos.

	Amplit.	Desvio abs	Variancia	D. Padrão	C. Var.
Bush	182.996	2.239.468	2.185.244.560	46.747	123
Buchanan	1.001	10.861	49.000	221	108

Quando o efeito da maior quantidade de votos é retirado, isto é, na análise do coeficiente de variação observa-se que a votação do candidato Bush tem uma dispersão mediana enquanto a do candidato Buchanan é bastante mais homogênea.

Assumindo que haja uma correlação entre os votos dos dois candidatos é possível realizar uma regressão linear a partir dos dados dos votos dos condados, como mostra a figura abaixo.

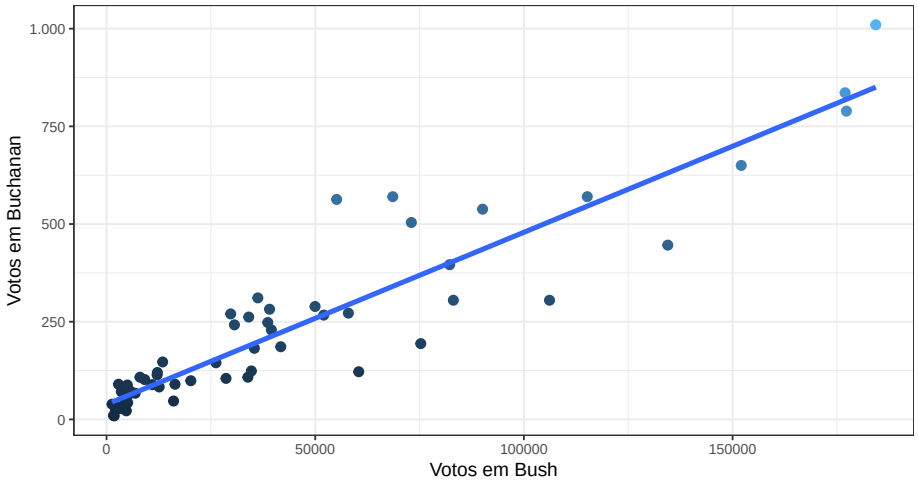
Figura 9: Relação entre votos recebidos por Gore e por Bachanan nos condados da Flórida. (a) Dados completos (Gore), (b) Sem dados discrepantes (Gore), (c) Dados completos (Total), (d) Sem dados discrepantes (Total)



Fonte: 2000 OFFICIAL PRESIDENTIAL GENERAL ELECTION RESULTS – USA

A Figura 10 mostra a reta de regressão linear obtida, que tem como função $f(x)$ = -2447 + 196.39x. Com base nesta regressão é possível inferir que caso o candidato Bush tivesse recebido 152.846 votos em um determinado condado, a votação do canditato Buchanan seria de aproximandamente 791 votos.

Figura 10: Relação entre votos recebidos por Bush e por Bachanan nos condados da Flórida e sua regressão linear.



É possível elaborar um código de programação em linguagem R que realize esta mesma previsão utilizando-se do método de Estimativas de Mínimos Quadrados, para se obter $\hat{\beta}_0$ e $\hat{\beta}_1$ e assim, prever um valor de votação do canditato Buchanan sob as mesmas condições, como se segue:

```
x = dados$BUSH
y = dados$BUCHANAN
n = length(x)
beta1 = (sum(x*y) - as.numeric(sum(x)) * as.numeric(sum(y))/n) /
  (sum(x^2) - ((sum (x))^2)/n)
beta0 = mean(y) - beta1 * mean(x)
votos = round(beta0 + 152846 * beta1, 0)

cat("Caso o candidato Bush tivesse recebido 152.846 votos em um
    determinado condado, a votação do canditato Buchanan seria
    de aproximandamente", votos," votos.")
```

Executando o código acima, obtém-se a estimação por mínimos quadrados de 797 votos. Observa-se que por regressão linear a resposta foi de 791 votos, desta forma, percebe-se que ambos os métodos foram similares na predição da votação do candidato Buchanan.

Conclusão

Diante do modelo gerado baseado nas 104 características coletadas dos gambás se conclui que a cada unidade de comprimento acrescida a largura do crânio cresce 0,36 unidades de comprimento e na condição hipotética de se identificar um comprimento nulo, a largura do crânio seria de 25 unidades de comprimento.

Mesmo que em certas condições, existam afirmações impossíveis de serem sustentadas, o que acertadamente se constata é que a largura do crânio dos gambas é diretamente proporcional ao comprimento total do mesmo.

Quanto a análise do dados da eleição presidencial de 2000 nos Estados Unidos, a atividade permite avaliar um modelos de predição que tem boa aderência aos dados disponibilizados, ao mesmo tempo que mostra a fragilidade da técnica quando assume que a correlação entre as votações dos candidatos pode ser explicada apenas pelos votos observados.