# Projeto 1

Gabriel Victor Soares dos Santos RA:197563

01/10/2021

## 1. Introdução

Este trabalho tem como objetivo a aplicação de técnicas estatísticas fazendo análises descritivas, exploratórias e inferenciais sobre um modelo de regressão linear simples. Além de servir como método avaliativo para a matéria **ME613** - Análise de Regressão da Universidade Estadual de Campinas.

Dentre os bancos de dados disponíveis, foi escolhido "*teengamb*", um conjunto de dados coletados em 1988, pelo departamento de psicologia da Universidade de Exeter, Inglaterra, que pesquisava sobre a menor idade de apostas em adolescentes britânicos.

## 2. Metodologia

Os pesquisadores, Susan G. Ide-Smith e Stephen E. G. Lea, fizeram um questionário de 9 páginas, com uma amostra de 51 adolescentes, com média de 13.7 anos, em uma escola de Exerter. Desse questionário, estão presentes neste banco de dados as seguintes variáveis: gênero, status socio-econômico de um dos pais, renda semanal do adolescente (em libras), gasto anual em apostas (em libras) e *verbal* que media a inteligência pela Escala de Vocabulário Mill Hill (*Mill Hill Vocabulary Scale, abreviadamente MHV*).

Neste trabalho é utilizada a linguagem de programação R e o programa *RStudio* para os testes, cálculos e criação de tabelas e gráficos. O refinamento dos modelos encontrados não serão feitos, pois esse conteúdo não faz parte da matéria dada até a entrega deste trabalho.

*Obs: Não foram encontradas informações suficientes a respeito da classificação da variável "status" desse conjunto de dados e será tratada como uma variável quantitativa*

## 3. Descrição Dados

O conjunto é composto de 5 variáveis, com 47 observações ao todo porque 4 questionários foram descartados pelos pesquisadores por não conter informações suficientes.

A Tabela 1 apresenta o número de participantes de cada gênero e a Tabela 2 algumas medidas sumárias da amostra inteira. A renda semanal foi multiplicada pelo número de semanas em um ano, a fim de facilitar cálculos e criar uma equivalência com o gasto anual em aposta.

Table 1: Total de Participantes por Gênero

| Gênero | Participantes |
|---|---|
| Feminino | 28 |
| Masculino | 19 |

Table 2: Medidas Sumárias das Variáveis Quantitativas

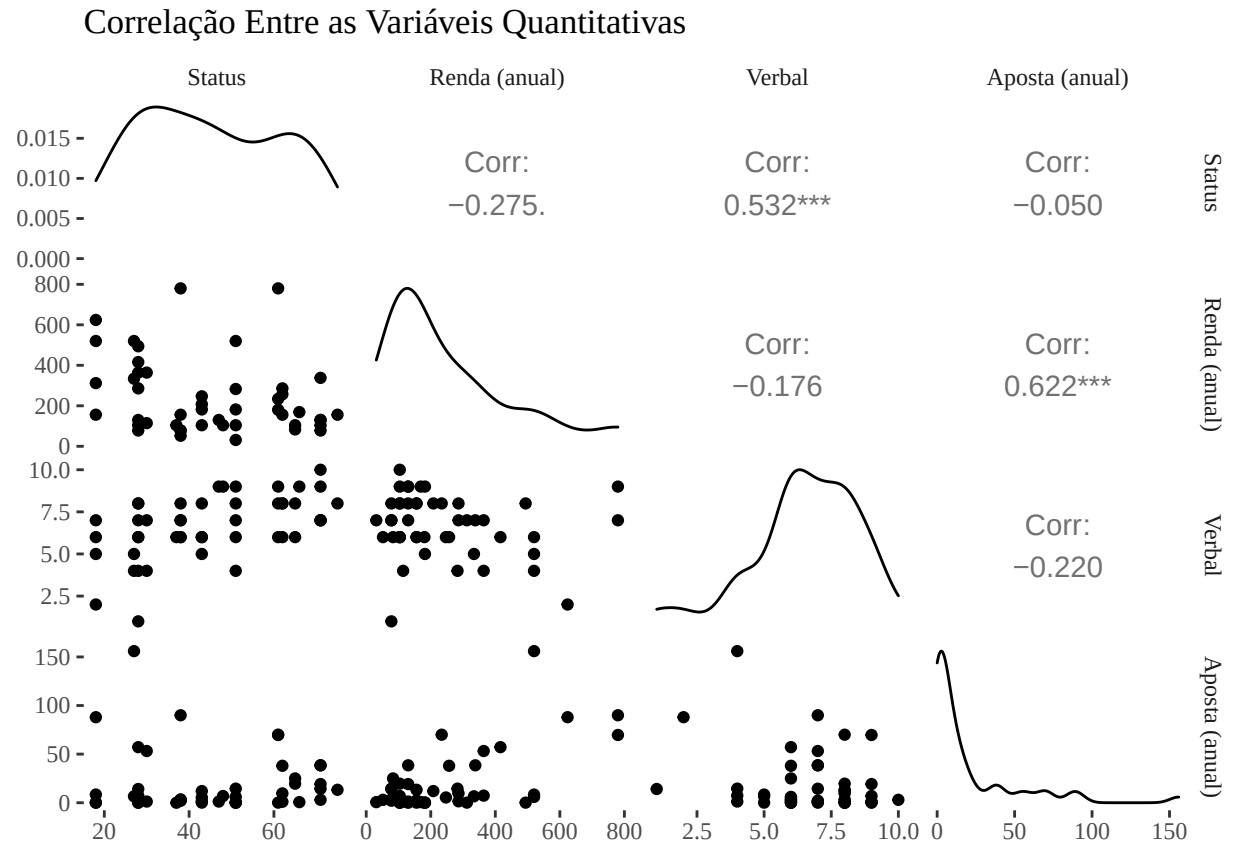
| | Média | Desvio Padrão | Mediana | Valor Mínimo | Valor Máximo |
|---|---|---|---|---|---|
| Renda (anual) | 241.379574 | 184.671304 | 169 | 31.2 | 780 |
| Status | 45.234043 | 17.262944 | 43 | 18.0 | 75 |
| Verbal | 6.659574 | 1.856558 | 7 | 1.0 | 10 |
| Gamble | 19.301064 | 31.515866 | 6 | 0.0 | 156 |

Observando a mediana do gasto de apostas anual vemos que é de 6 libras, enquanto a média é de quase 20 libras, o que poderia indicar uma baixa frequência em apostas nesta amostra, mas com participantes que apostam somas consideráveis de dinheiro.

A maioria dos participantes se concentra entorno de 7 pontos na Escala de Vocabulário Mill Hill, na variável *Verbal*, podendo indicar uma inteligência mediana da amostra.

# 4. Analise exploratória

Pela Correlação de Pearson é possível quantificar a correlação entre as variáveis, além de determinar se a variáveis são direta, ou inversamente, proporcionais




Segundo o gráfico a menor correlação foi de -22%, entre a variável resposta e a variável *Verbal*, que é considerado uma correlação negativa fraca. A maior correlação foi de 62,2% entre a variável Renda e Aposta, sendo uma correlação moderadamente positiva. Quanto a correlação entre Aposta e Status é quase inexistente por ser próximo de 0.

Destaca-se a seguir os gráficos de dispersão mais importantes, eles pode nos indicar visualmente um pouco do que foi apresentado no gráfico de Correlação de Pearson, apresentados anteriormente.

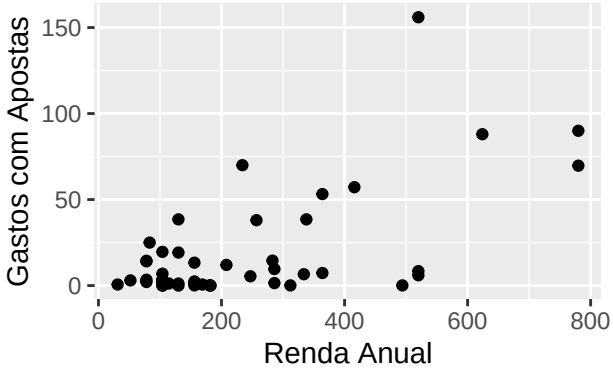



Gráfico 1

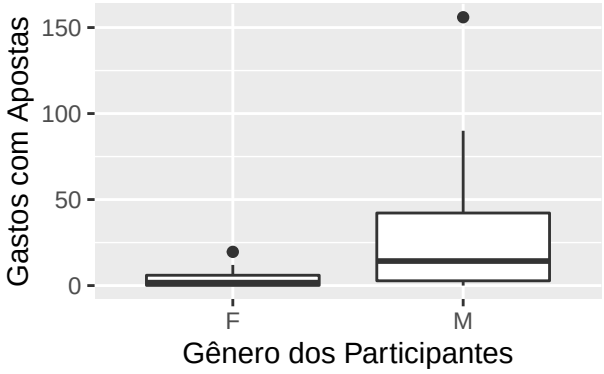



Gráfico 2

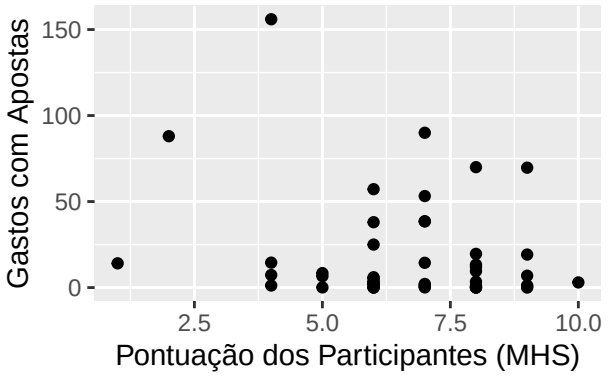



Gráfico 3

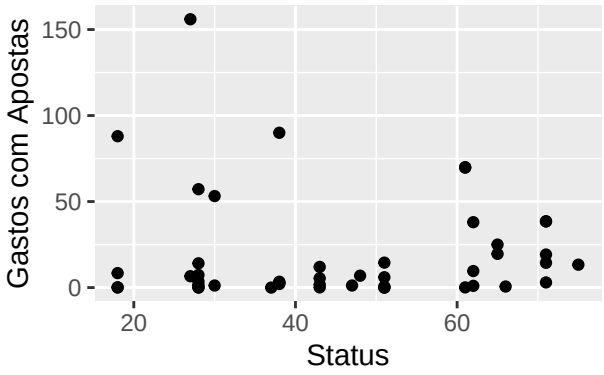



Gráfico 4

Baseado na nossa amostra e nesses gráficos, ainda que exista correlação entre as variáveis, os dados não aparentam visualmente ser muito lineares.

O gráfico 1 mostra uma concentração de dados quando a renda anual é inferior a 200 libras, como indicado pela mediana da tabela 2. O gráfico 2 mostra que os participantes do gênero masculino apostam mais anualmente que o gênero feminino. Assim como na tabela 2, o gráfico 3 mostra melhor que a distribuição dos participantes, em relação a Escala de Vocabulário Mill Hill, centrada entorno do 7.

## 5. Análise Inferencial

O modelo de regressão utilizado nesse trabalho será:

$$Y_i = \beta_0 + \beta_1 X_1 \tag{1}$$

, onde:

- $\beta_0, \beta_1$ são parâmetros.
- $X$ é a constante conhecida.
- $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ é um termo de erro aleatório.
- $i = 1, 2, ..., n$.

Serão construídos modelos de regressão simples entre a variável resposta, neste caso Gasto Anual em Apostas, e as demais variáveis, apresentado a equação da reta, o gráfico da regressão e os modelos encontrados serão testados a fim de determinar se são adequados.

Avalia-se primeiro a Análise de Variância (ANOVA) para encontrar se há diferença entre a distribuição das variáveis. Se o modelo for adequado, verifica-se então a normalidade pelo Teste de Shapiro-Wilks, heterocedasticidade pelo Teste de Breuch-Pagan e linearidade por um gráfico de resíduos.

## 5.1 Status e Apostas

A equação do modelo é:

$$Y_i = 23.46 - 0.092X_1 \tag{2}$$

E o gráfico de regressão

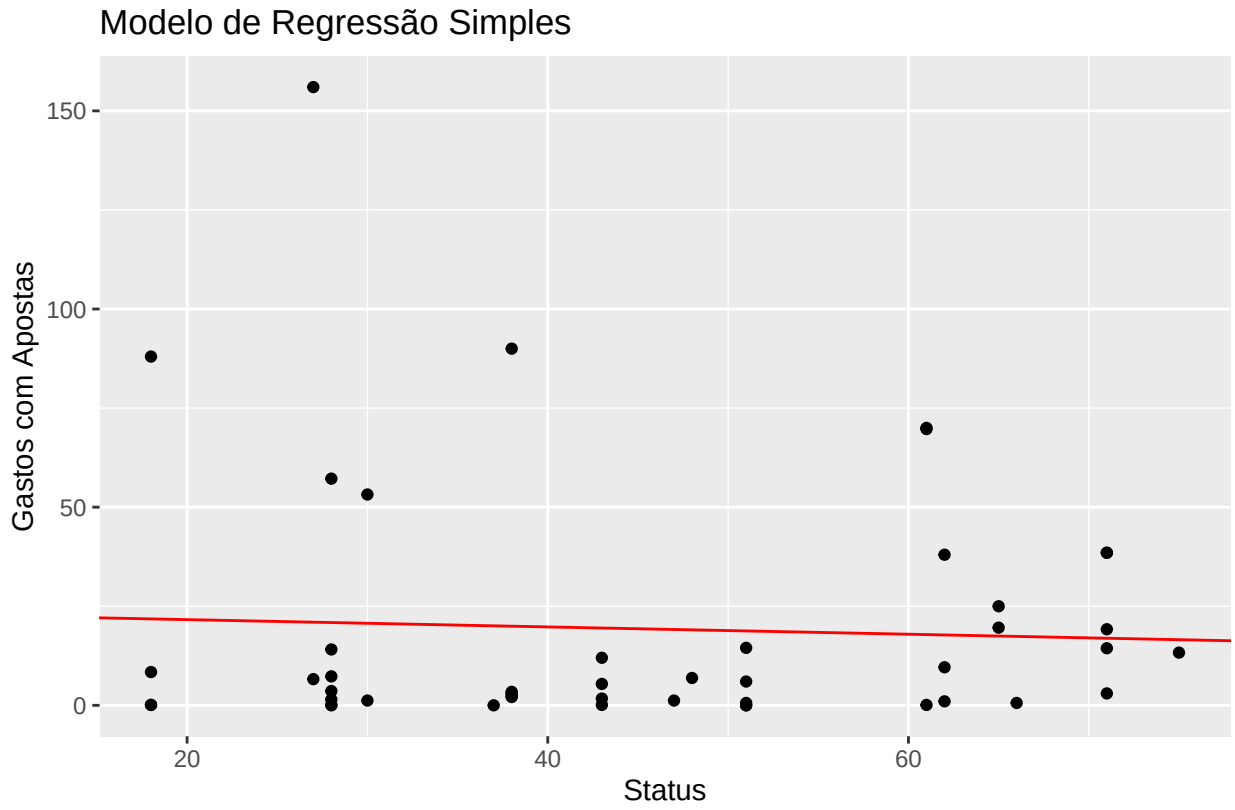



Gráfico 5

| Teste ANOVA | p-valor | Resultado |
|---|---|---|
| Modelo Status | 0.7364 | Fracasso |

Entretanto, o p-valor é superior a 0.05. Portanto não é um modelo adequado.

## 5.2 *Verbal* e Apostas

A equação do modelo é:

$$Y_i = 44.17 - 3.73X_1 \tag{3}$$

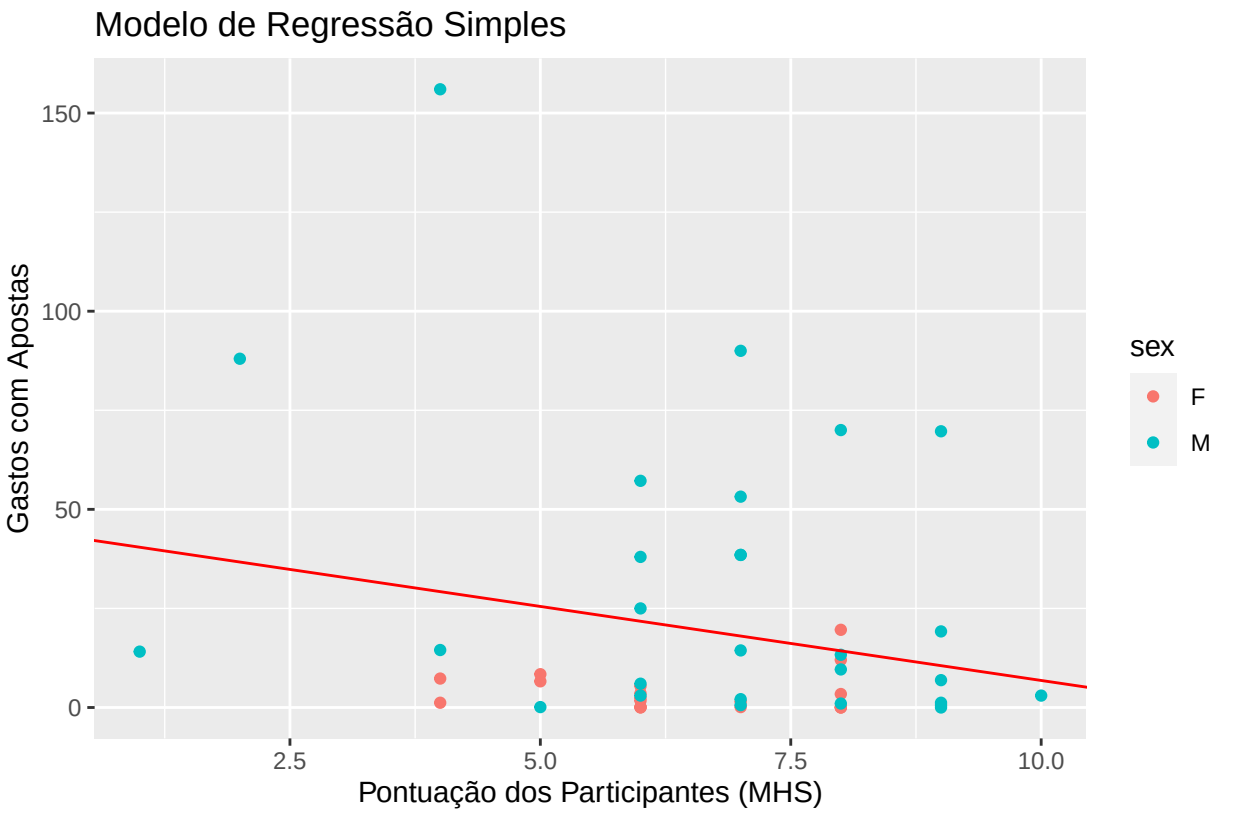



Gráfico 6

| Teste ANOVA | p-valor | Resultado |
|---|---|---|
| Modelo Verbal | 0.1372 | Fracasso |

Entretanto, o p-valor é superior a 0.05. Portanto não é um modelo adequado.

### 5.3 Gênero e Apostas

A equação do modelo é:

$$Y_i = 3.86 + 25.90X_1 \tag{4}$$

| Teste ANOVA | p-valor | Resultado |
|---|---|---|
| Modelo Gênero | 0.004437 | Sucesso |

O p-valor do teste desse modelo é inferior a 0.05. Portanto não há diferença entre a distribuição dessas variáveis.

## 5.4 Renda e Apostas

Baseado no contexto dos dados, não é possível ter renda negativa, seria rasoável supor que 0 de renda deve se relacionar a 0 gastos em apostas, se considerar que os participantes com 0 de renda não apostem com dinheiro alheio. Este modelo de regressão linear será feito pelo ponto de origem (0,0).

Este modelo apresenta a seguinte equação:

$$Y_i = 0.089X_1 \tag{5}$$

E seu gráfico segue abaixo:

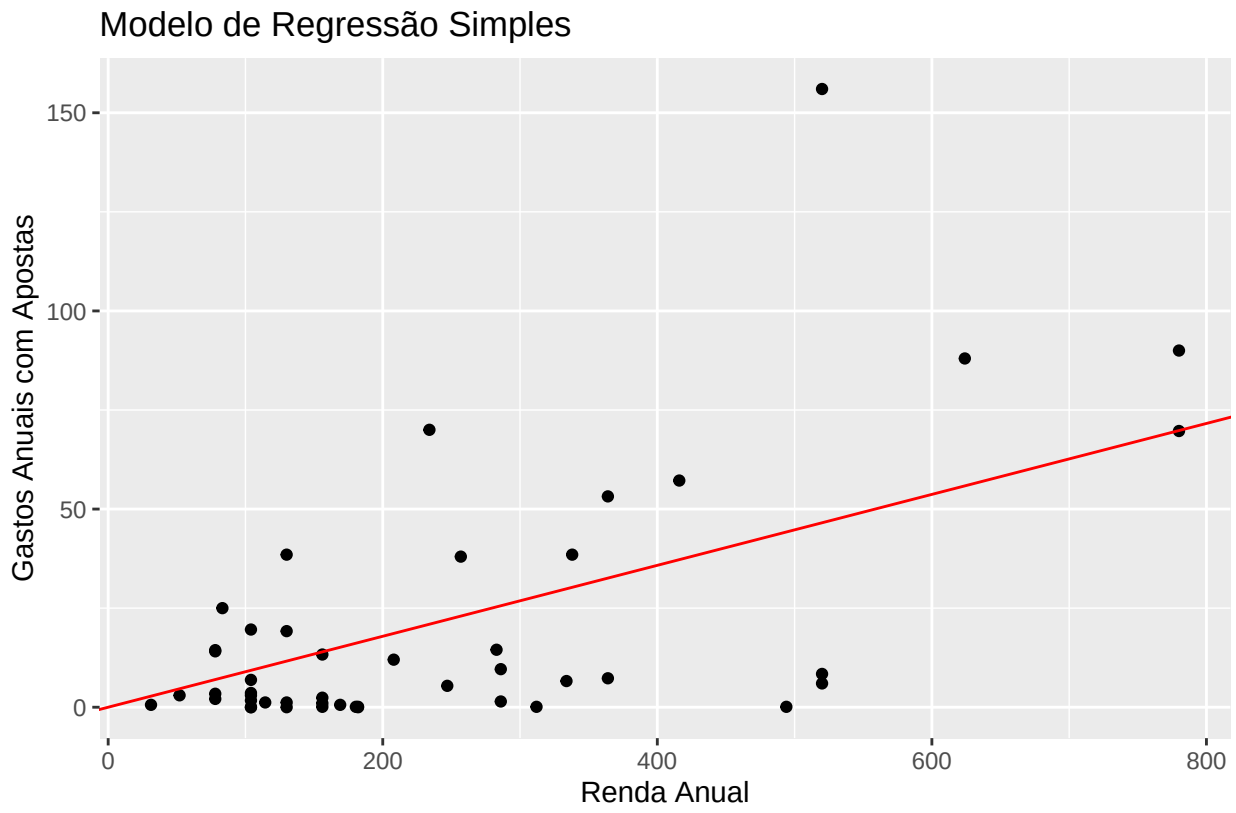



Gráfico 8

| Teste ANOVA | p-valor | Resultado |
|---|---|---|
| Modelo Renda | 2e-07 | Sucesso |

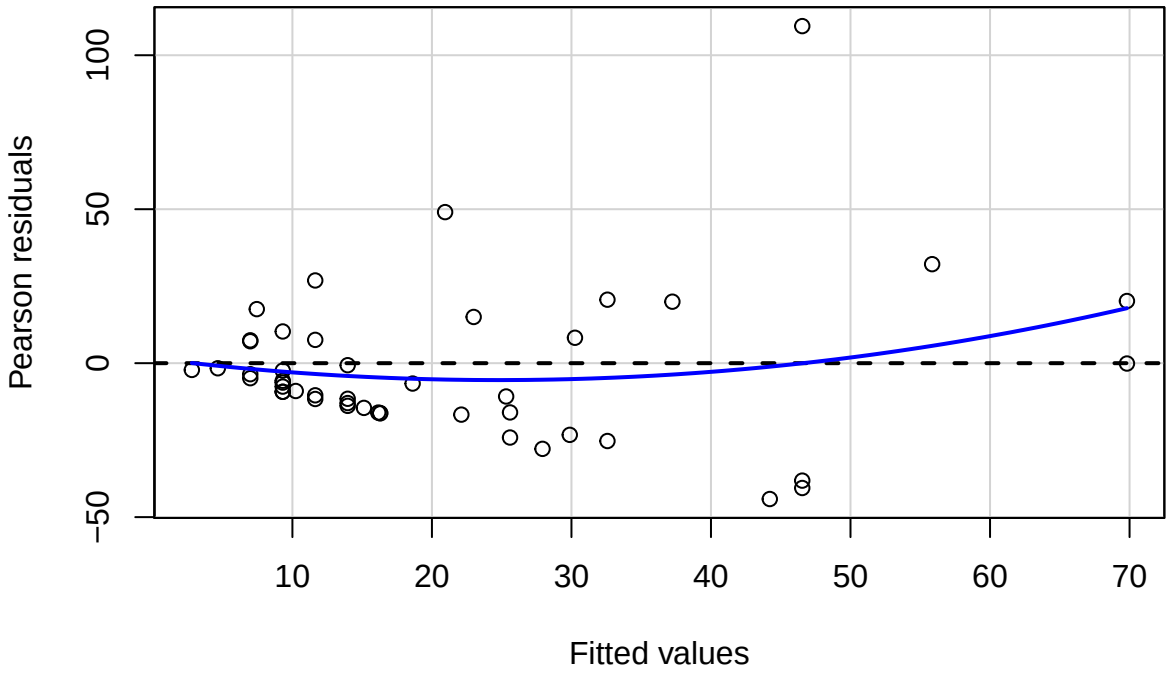
o p-valor do teste desse modelo é inferior a 0.05. Portanto não há diferença entre a distribuição dessas variáveis.

Dos 4 modelos encontrados, apenas 2 foram considerados adequados pelo teste de Análise de Variância.

### 5.5 Linearidade e Erros Independentes

O gráfico de resíduos permite observar a linearidade do modelo e independência das variáveis. Portanto, espera-se ver uma distribuição normal com média 0 ao longo do eixo y, e os erros resíduais estarem distribuídos mais igualitariamente ao longo do eixo x.

Os resíduos são, de certo modo, a diferença da média com os valores observáveis, como o modelo para gênero é dicotômico, seu gráfico de resíduo



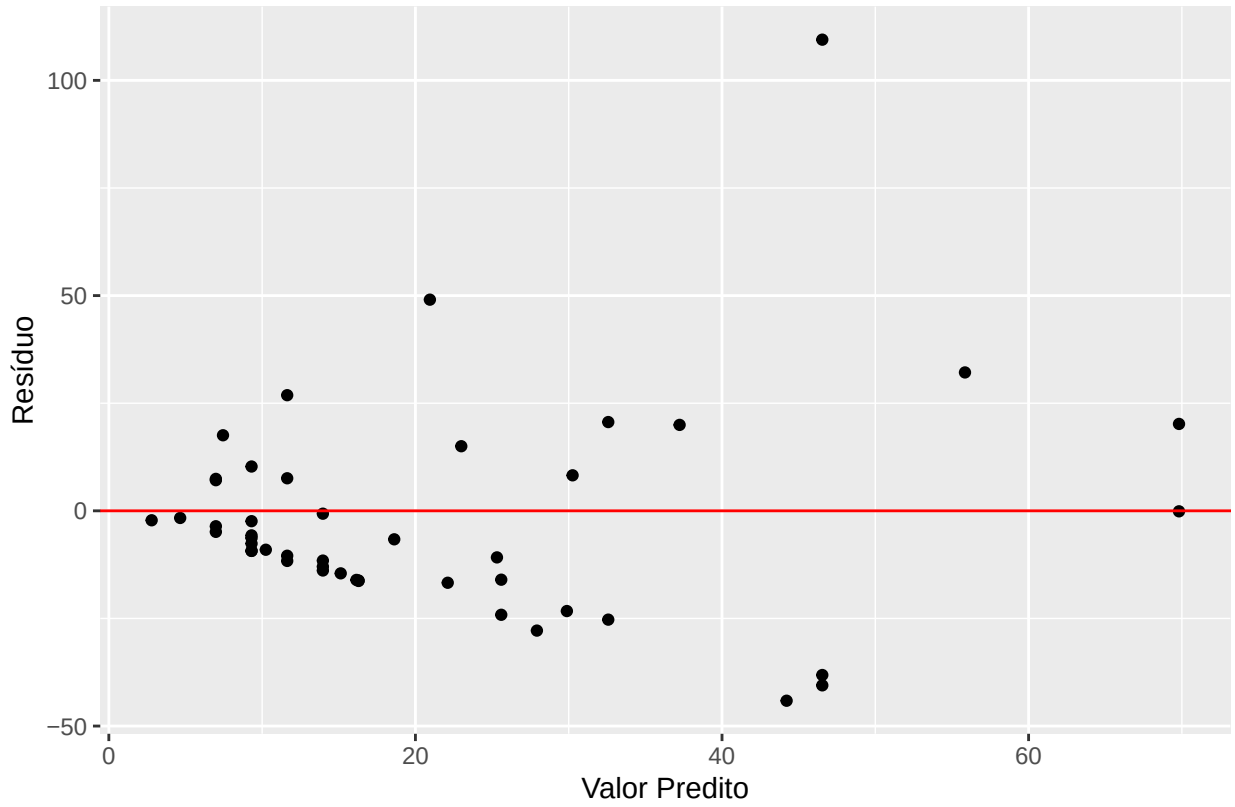



Gráfico 9

O gráfico de resíduos não tem uma variância na distribuição dos pontos, apontando dependência entre as variáveis Renda e Apostas. Tem alguma distribuição entorno de 0, entretanto será usado o teste de Shapiro-Wilks para determinar se há normalidade no modelo. O gráfico indica que um modelo não linear pode ser mais adequado,

### 5.6 Teste de Shapiro-Wilks e Teste de Breuch-Pagan

Para a avaliação do modelo serão realizados dois tipos de diagnósticoso: teste de Shapiro-Wilks, que busca analisar a normalidade dos resíduos e o teste de Breuch-Pagan, que, supondo normalidade, avalia a heterocedasticidade. O critério de decisão em ambos os testes, se o p-valor for inferior a 0.05, o modelo será considerado um sucesso.

| Teste de Shapiro-Wilks | p-valor | Resultado |
|---|---|---|
| Modelo Renda | 1.59e-05 | Fracasso |
| Modelo Gênero | 1.20e-06 | Fracasso |

O p-valor para o teste de Shapiro-Wilks, de ambos os modelos, é inferior a 0.05. Portanto, não tem distribuição normal, e mesmo não que o modelo não siga as condições de normalidade, será avaliado o teste de Breuch-Pagan.

| Teste de Breusch-Pagan | p-valor | Resultado |
|---|---|---|
| Modelo Renda | 0.0000010 | Fracasso |
| Modelo Gênero | 0.0001091 | Fracasso |

O p-valor para o teste de Breuch-Pagan, de ambos os modelos, é inferior a 0.05. Portanto, os dados não tem uma distribuição regular.

# 6 Conclusão

Baseado na amostra de dados e nos testes realizados, existe uma frequência maior dos participantes do gênero masculino que apostam e uma correlação moderada entre renda e gasto em apostas pode indicar que quanto maior for a renda, mais gastaria em apostas.

Os modelos de regressão encontrados não foram considerados adequados e é necessário refinamento do conjunto para montar um modelo mais apropriado.

# 7. Bibliografia

IDE-SMITH S. G., Lea S. E. G., (1988). Journal of Gambling Behavior, cap.4, pgs.110-118. Disponível em: "http://www.utstat.utoronto.ca/reid/sta2201s/2012/teengamb.pdf"

KUTNER, M. H., NACHTSHEIM, C., NETER, J., & LI, W. (2005). Applied Linear Statistical Models.