Relatório #1 Análise de Regressão

Regressão linear baseado no total mensal de passageiros transportados no Metrô no Município do Rio de Janeiro entre 1998 e 2023

José Victor Pereira Fuks – 119021632

Vinícius Rabello Rodrigues - 119056899

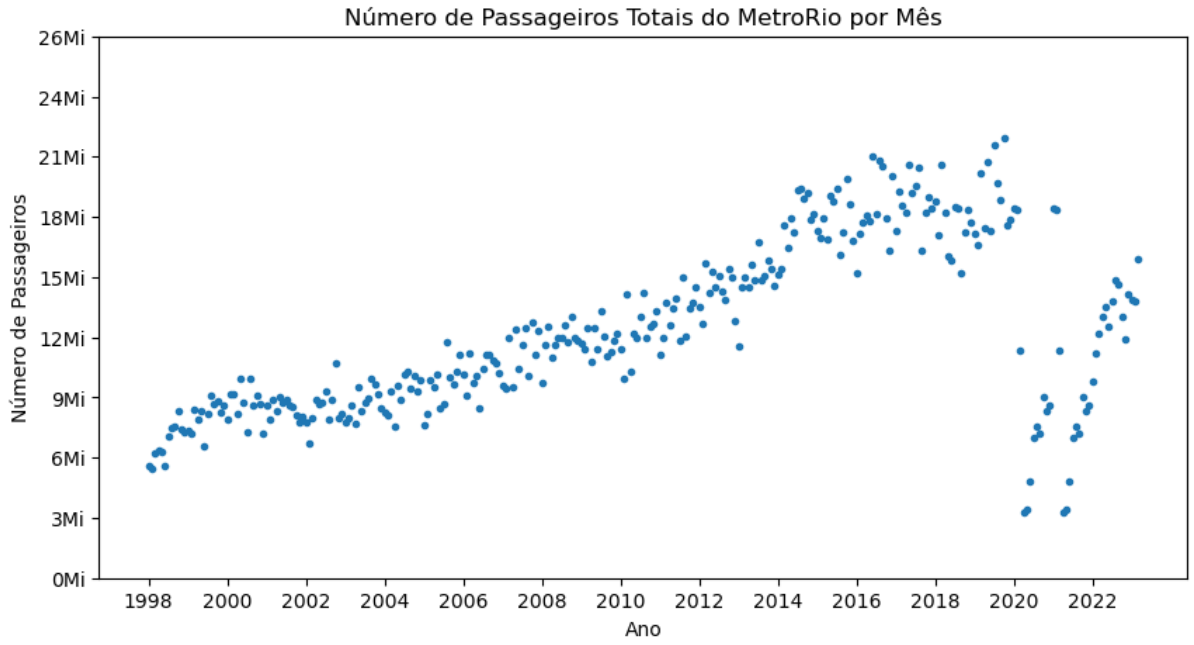
Nicholas Morrey-Jones Santos – 119062434

Introdução

Este relatório apresenta uma análise dos dados do MetrôRio, com o número de passageiros por mês, em milhares, do ano de 1998 até 2023. Os dados são fornecidos pelo próprio governo do Rio de Janeiro, através do instituto Pereira Passos, os dados vem no formato de tabelas do excel, uma para cada ano, tabelas em que cada linha é uma estação, e as colunas o total mensal de passageiros por mês.

Nosso objetivo é estudar a evolução deste meio de transporte público ao longo dos anos, os efeitos da pandemia no mesmo, e finalmente, prever o número de passageiros dos próximos meses, visto que é importante saber a demanda futura deste serviço, para que sejam feitos os devidos investimentos neste setor do transporte público. Portanto, nossa covariável será o tempo (meses) e a variável resposta será o número de passageiros. Como não nos interessa as linhas e estações, transformamos os dados fornecidos em um data frame com X = meses e Y = passageiros totais do metrô.

Abaixo está o gráfico do Número de Passageiros Totais do MetrôRio por Ano, cada ponto representa um mês:

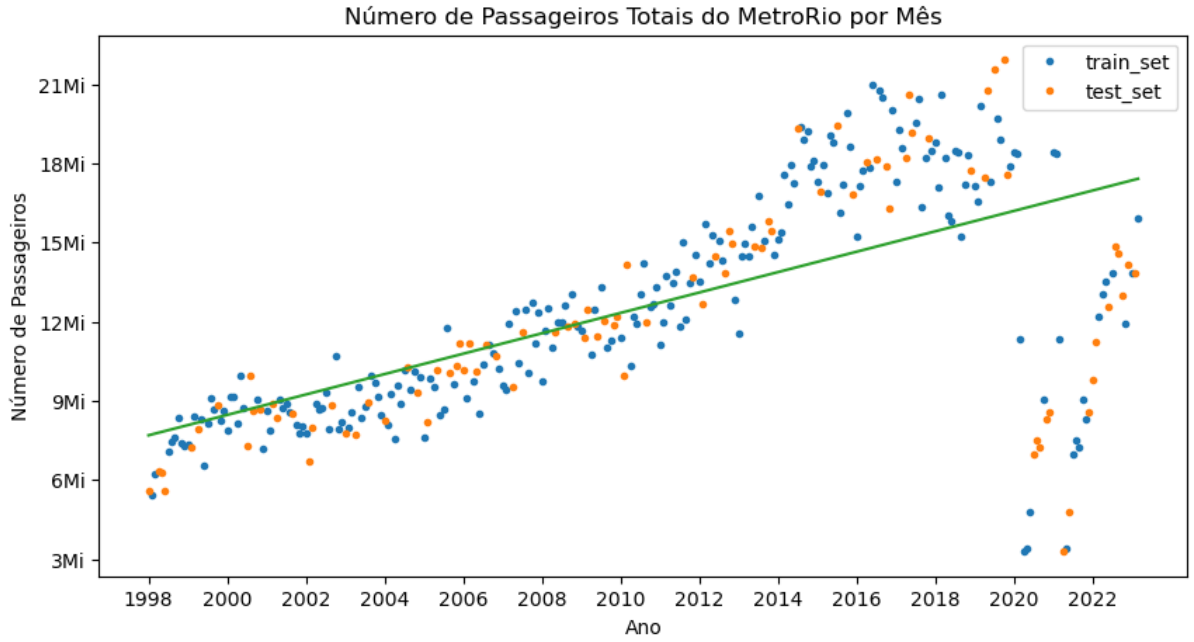


Note que parece haver uma relação linear entre Y e X até o ano de 2020, quando começa a pandemia de Covid-19 no Rio de Janeiro, e há uma queda brusca no número de passageiros. Vale notar também que o número cresce durante 2020 até cair novamente, de forma similar no começo de 2021, para voltar a aumentar durante o mesmo ano e até os dias de hoje. Perceba que mesmo com o fim da pandemia, o número de passageiros é menor do que era antes dela, um dos motivos pode ser a popularização do home office, acabando com a necessidade de parcela da população de usar o Metrô como meio de locomoção até o trabalho.

Primeiro fazemos um modelo para os dados de 1998 até 2023, dividindo nossos dados em conjuntos de treino e teste:

Caso 1 – Reta de regressão total mensal de passageiros transportados no Metrô no Município do Rio de Janeiro entre 1998 e 2023.

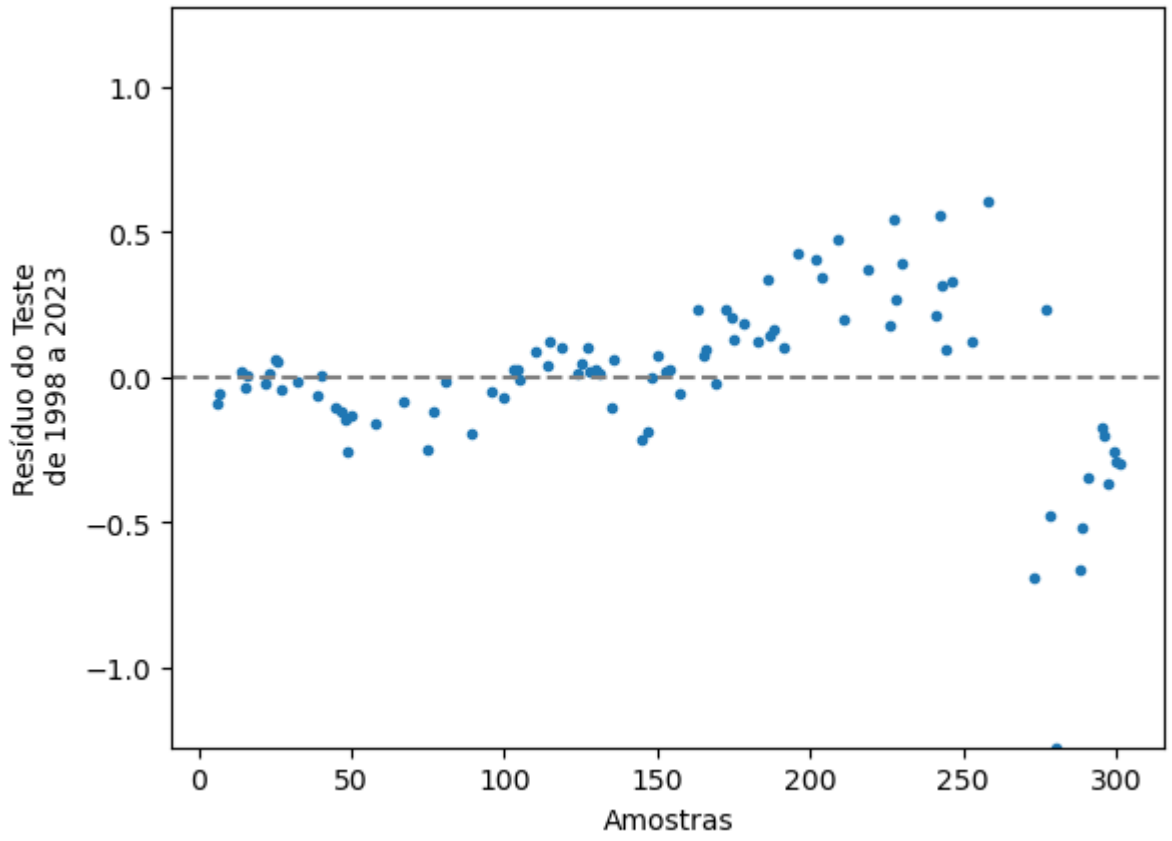
-Reta regressão por mínimos quadrados:




```
Residuals:
     Min       1Q   Median       3Q      Max
-1.28442 -0.10989 -0.01227  0.19780  0.65949

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.779129   0.048605   16.03   <2e-16 ***
train_98_23$x 0.002989   0.000280   10.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3568 on 210 degrees of freedom
Multiple R-squared:  0.3517,    Adjusted R-squared:  0.3487
F-statistic: 113.9 on 1 and 210 DF,  p-value: < 2.2e-16

 Analysis of Variance Table

 Response: train_98_19$x
               Df Sum Sq Mean Sq F value    Pr(>F)
train_98_19$y   1 953757  953757  1455.5 < 2.2e-16 ***
Residuals     182 119259     655
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Podemos ver que ambos coeficientes são significativos a 0.1% de forma que os dois têm importância o suficiente para a reta de regressão que não poderiam ser substituídos por zero sem queda considerável na qualidade do ajuste. R2 está em 35,17%, valor bem abaixo do ideal, certamente esse valor se dá pelas quedas durante a pandemia. Porém, apesar disso temos coeficientes bem significativos, vale a pena fazermos modelos antes e depois

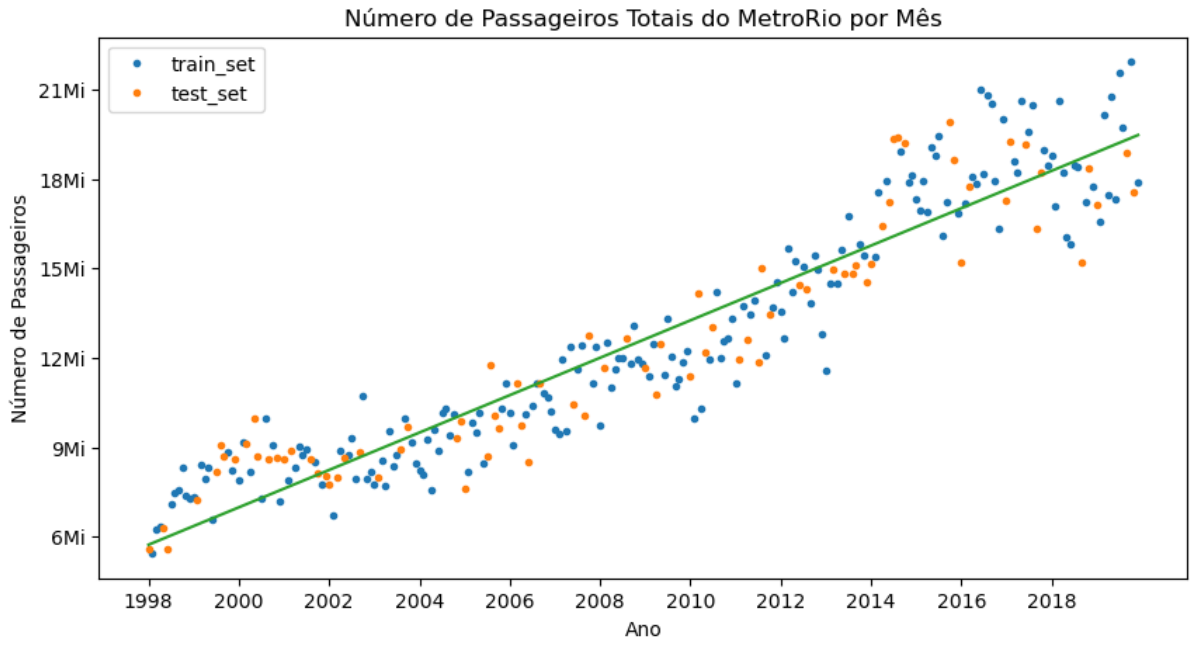
da pandemia, para estudarmos os resultados e possivelmente obtermos resultados melhores.



Vemos que os resíduos parecem no início sem correlação, até subir bastante e depois cair bruscamente, isso é por causa da inauguração da linha 4 do metrô em 2016 e da pandemia, respectivamente. Ainda assim os resíduos parecem não-correlacionados.

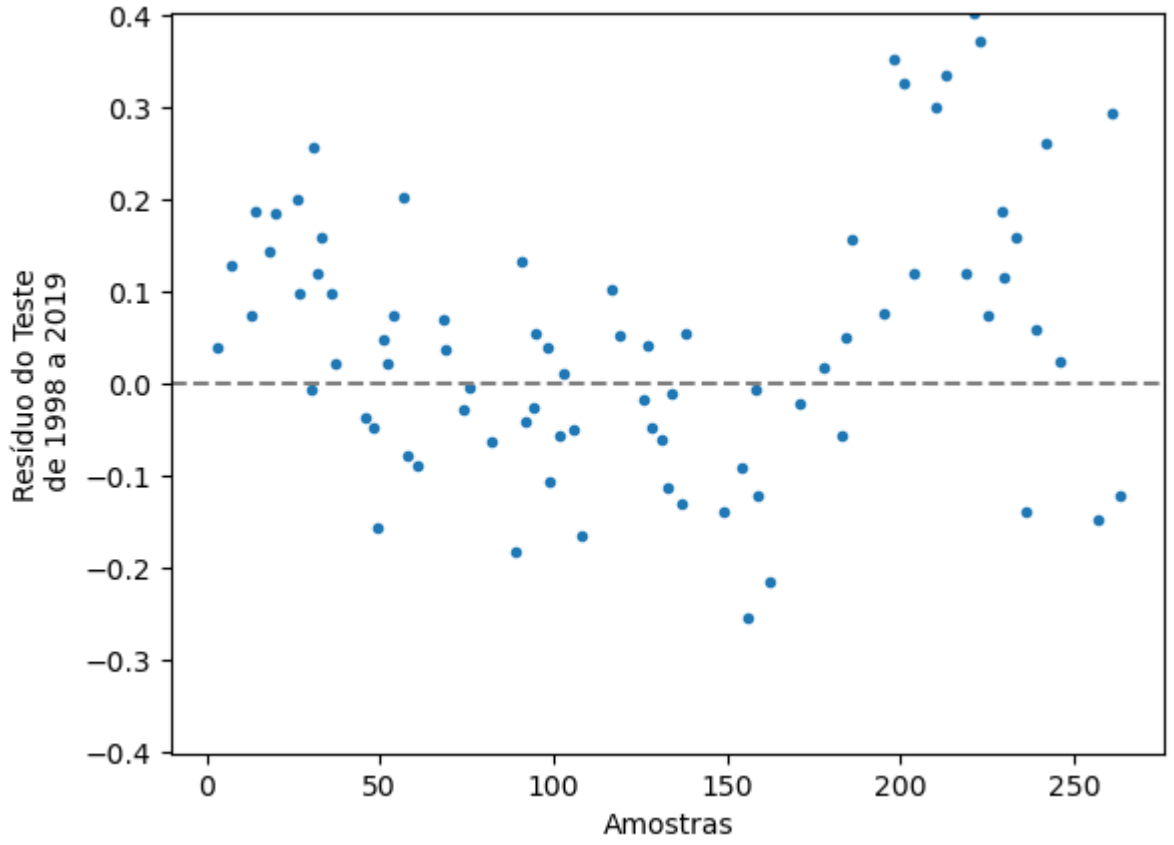
Caso 2 – Reta de regressão do total mensal de passageiros transportados no Metrô no Município do Rio de Janeiro entre 1998 e 2019 e entre 2020 e 2023 calculados separadamente.

-Reta 1998 a 2019 por mínimos quadrados:



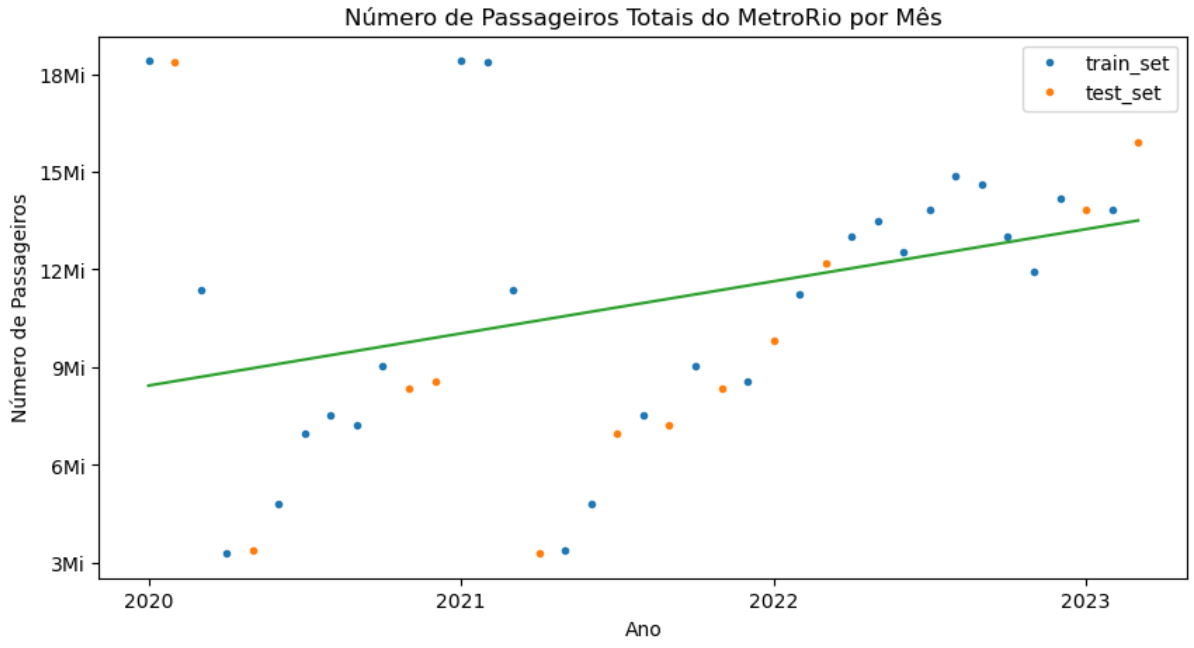

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.33386 -0.08684 -0.01186  0.08758  0.35268

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5824143  0.0204781   28.44   <2e-16 ***
train_98_19$x 0.0050445  0.0001322   38.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.137 on 182 degrees of freedom
Multiple R-squared:  0.8889,    Adjusted R-squared:  0.8882
F-statistic:  1456 on 1 and 182 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: train_98_19$y
              Df  Sum Sq Mean Sq F value    Pr(>F)
train_98_19$x   1 27.3049 27.3049  1455.5 < 2.2e-16 ***
Residuals     182  3.4142  0.0188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Novamente, obtemos coeficientes significativos a 0.1% mas agora temos um R^2 de 88%, se mostrando um modelo bem melhor que o primeiro, e mostrando que estávamos corretos em supor que a pandemia prejudicou bastante o ajuste anterior. Interessante ver que por mais que até 2019 fosse possível obter um excelente modelo linear, e que na época fizesse sentido o usar para prever o número de passageiros para os próximos meses, um evento externo como a pandemia pode mudar tudo.



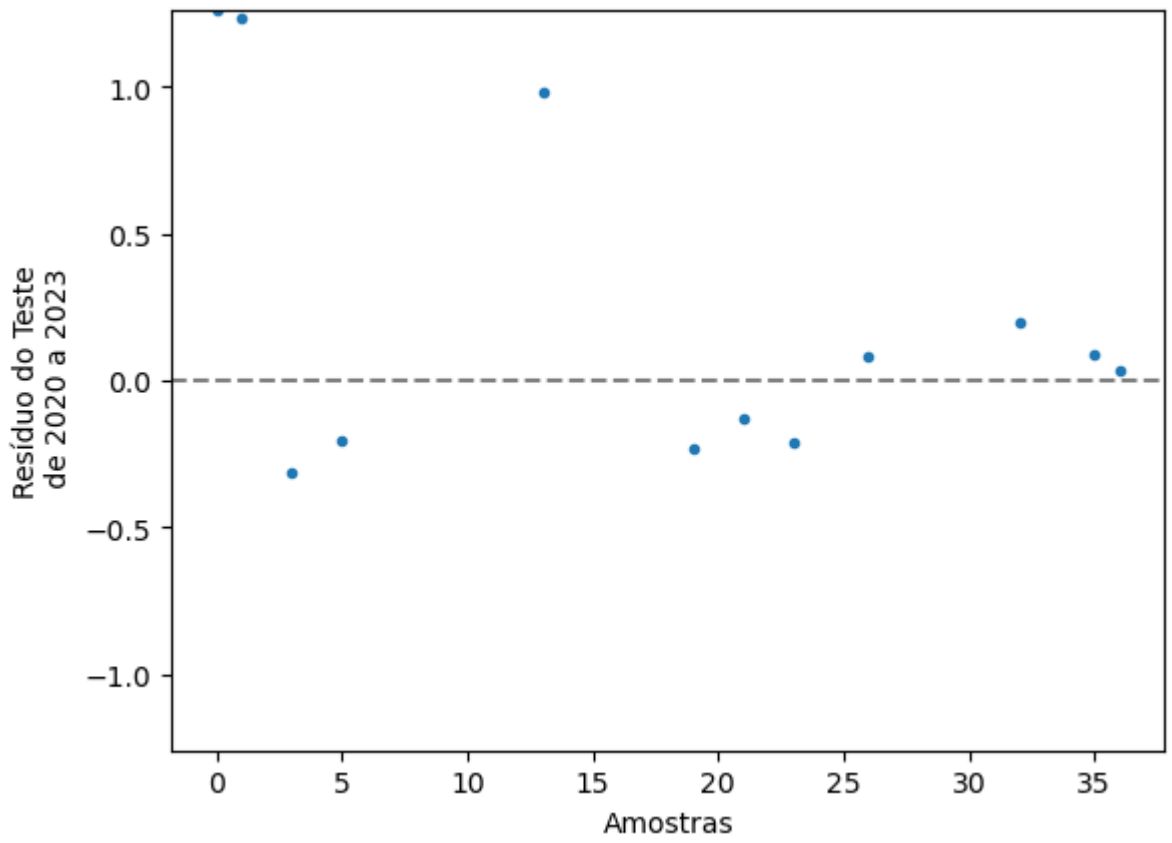
É possível ver um padrão que lembra uma função quadrática no gráfico dos resíduos, talvez utilizando uma transformação quadrática possamos deixar os resíduos mais próximos de não-correlacionados.

-Reta 20 a 23 por mínimos quadrados:



```
Residuals:
     Min       1Q   Median       3Q      Max
-0.58299 -0.16807  0.01366  0.15108  1.00268

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.580264   0.134679   4.308 0.000224 ***
train_20_23$x 0.021438   0.006077   3.528 0.001646 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3315 on 25 degrees of freedom
Multiple R-squared:  0.3324,    Adjusted R-squared:  0.3057
F-statistic: 12.45 on 1 and 25 DF,  p-value: 0.001646

Analysis of Variance Table

Response: train_20_23$y
              Df Sum Sq Mean Sq F value   Pr(>F)
train_20_23$x  1 1.3681 1.36810  12.447 0.001646 **
Residuals     25 2.7479 0.10992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nessa reta é perceptível pelo gráfico que o ajuste não é bom, o teste de significância para o coeficiente angular continua significativo a 0,1% mas fica em 1% para o coeficiente angular, ademais o valor do R^2 ficou em somente 33%, menor do que a reta do primeiro gráfico, assim como o valor do teste de significância da regressão que é significativo a 1% nesta reta. Essa queda na qualidade do ajuste está ligada à volatilidade dos valores dos dados, onde existe uma queda e um crescimento muito rápido de forma cíclica até 2022, onde o crescimento parece estar mais constante, por conta de políticas públicas na cidade do Rio de Janeiro sobre o Covid-19. Poderíamos ver também que os pontos mais altos estão na amostra de teste, o que influiu nos coeficientes da reta fazendo com que ela ficasse mais focada em possíveis outliers. Um modelo pós-pandemia seria mais interessante com mais tempo, no futuro, para termos uma amostragem maior, e talvez desconsiderando os anos de 2020 e 2021. Mesmo assim, parece haver um começo de um crescimento linear a partir do ano de 2021.



Os resíduos parecem não-correlacionados, tirando alguns outliers que são o começo de 2020 e o começo de 2021, antes de caírem como visto nos gráficos anteriores.

Conclusão

Como foi possível ver pelo gráfico 2, conseguimos de certa forma melhorar o ajuste em parte dos dados, mas com esse tipo de recorte temporal o período após a pandemia ficou com um ajuste pior do que a primeira reta. Por conta da volta de um período turbulento como o da pandemia e com dados ainda insuficientes para fazer uma boa previsão ou ajuste aos dados pós pandemia, para que ele fosse mais bem ajustado necessitaríamos de mais dados dos próximos meses ou anos para compensar a volatilidade do período da pandemia. Podemos ver que após a pandemia temos o começo de um crescimento linear do número de passageiros, mesmo que com menos passageiros que nos anos pré-Covid, talvez por causa do aumento da prática do trabalho semi-presencial ou remoto como foi dito, a tendência é aumentar. É possível que com o tempo o primeiro ajuste desempenhe melhor, com as coisas voltando ao normal, podemos ter que o número de passageiros volte ao que era antes, com uma taxa de crescimento parecida, mas precisamos de mais tempo e dados para fazer esse tipo de afirmações.