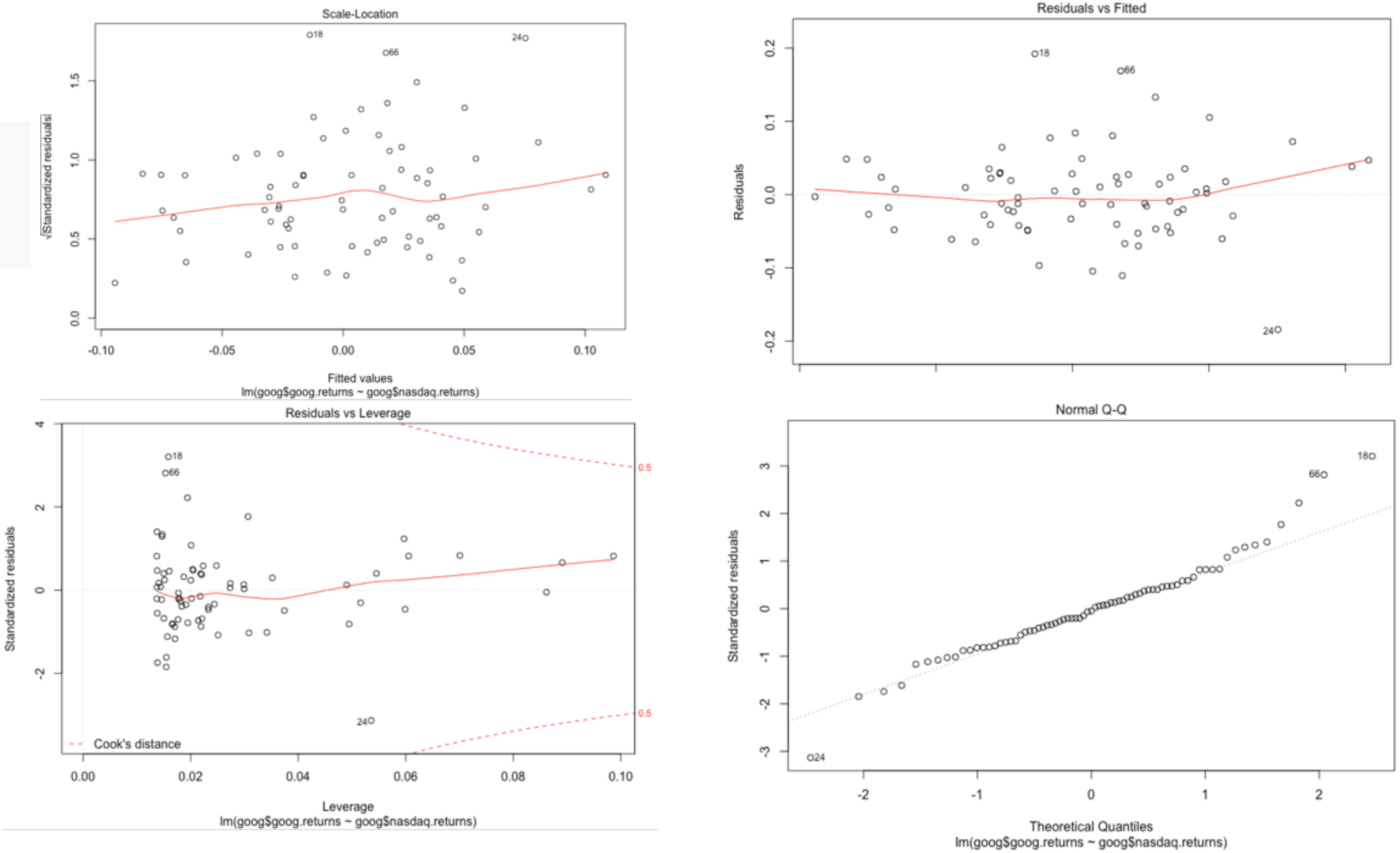
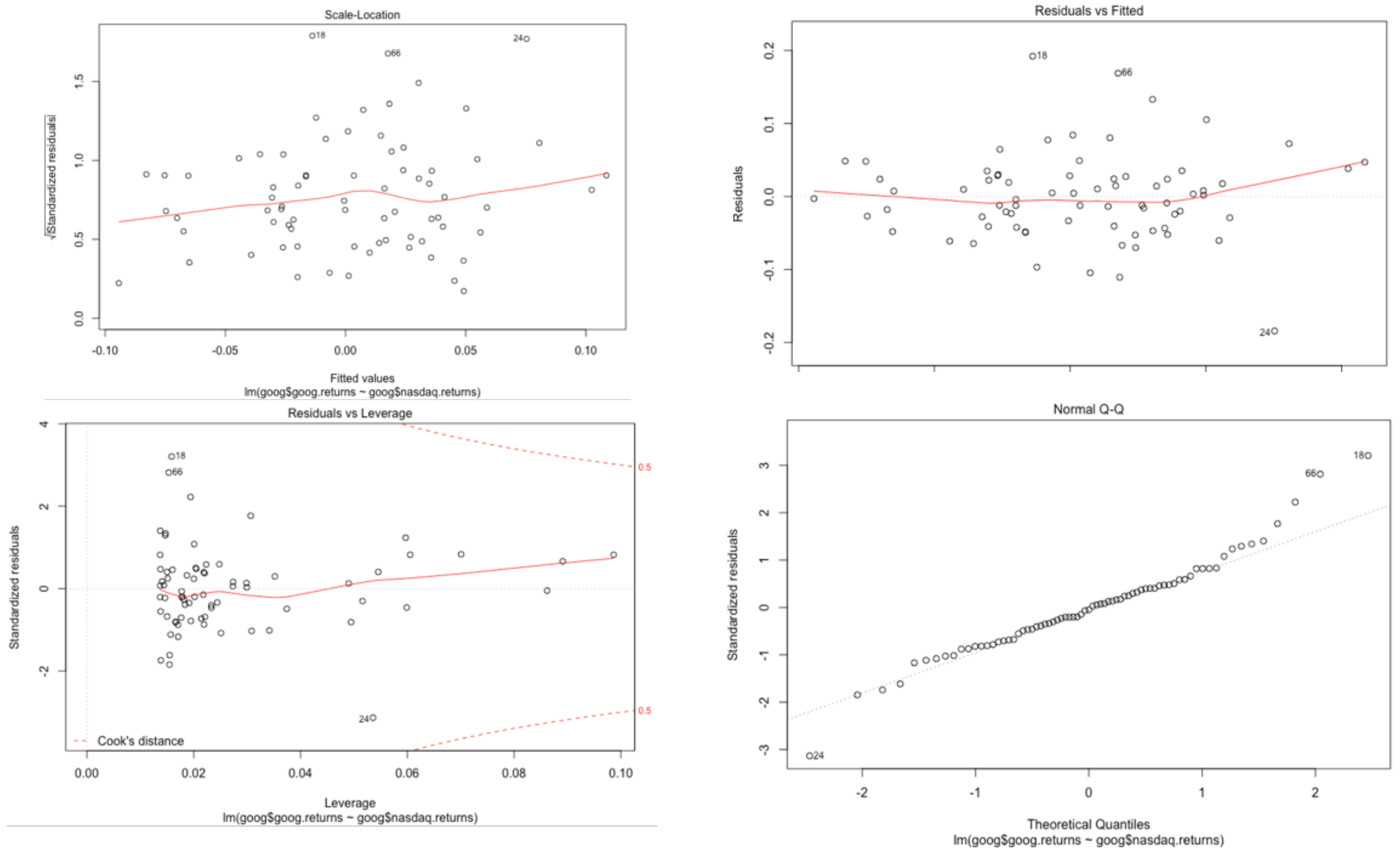
EXAMPLE 6: PARSING RESIDUAL PLOTS

IF YOU USE THE PLOT FUNCTION ON OUR LINEAR REGRESSION MODEL, IT PRINTS A BUNCH OF DIAGNOSTIC PLOTS OF THE RESIDUALS

```
plot(googM)
```



LINEAR REGRESSION IS VALID ONLY UNDER CERTAIN ASSUMPTIONS



THESE PLOTS HELP US CHECK WHETHER OUR DATA VIOLATES THESE ASSUMPTIONS

ASSUMPTION 1:

THE RESIDUALS ARE NORMALLY DISTRIBUTED

ASSUMPTION 1:

THE RESIDUALS ARE NORMALLY DISTRIBUTED

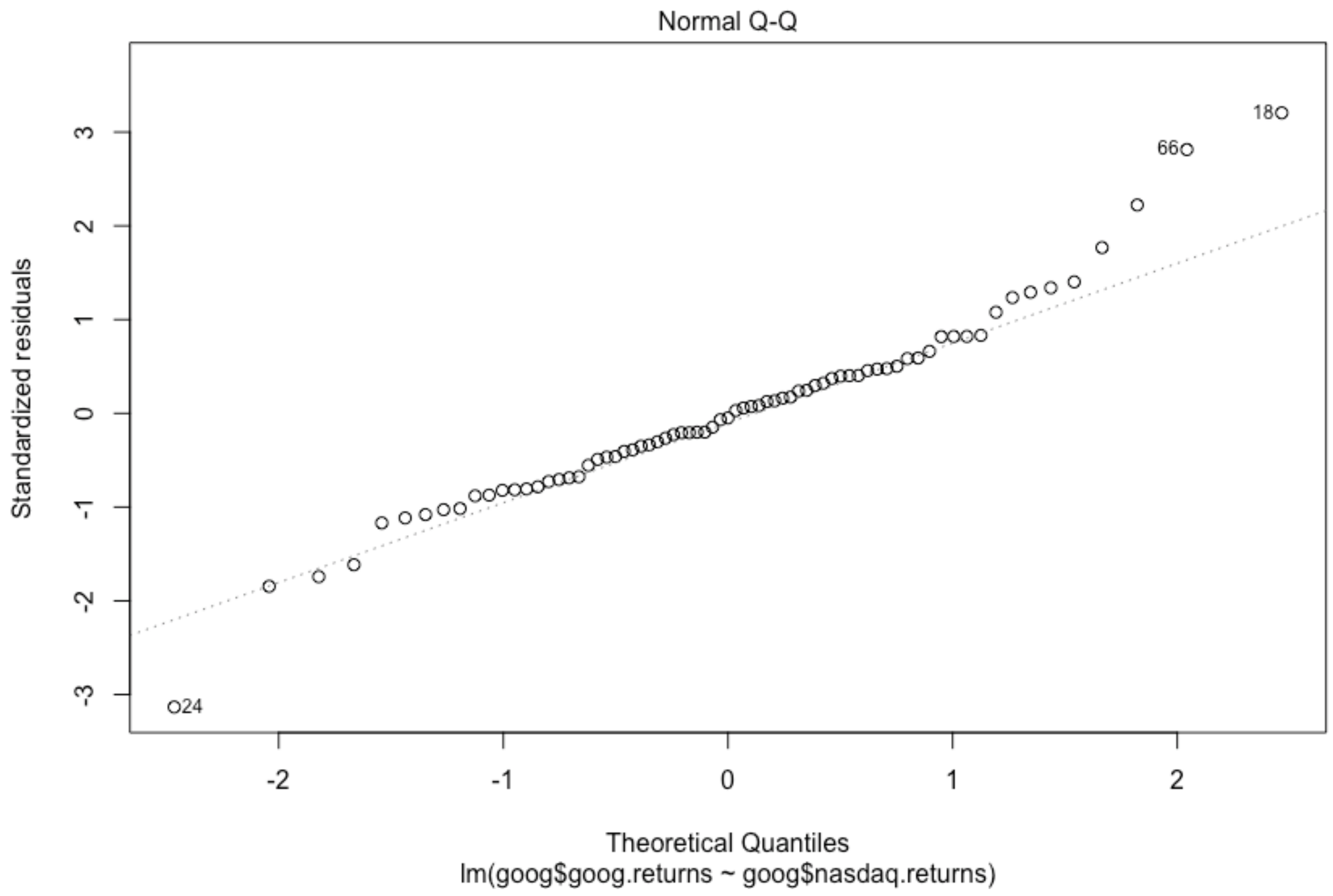
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.167102 -0.027855  0.004201  0.034741  0.121227

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.001029   0.022894  -0.045   0.9643
goog$nasdaq.returns  0.810490   0.163540   4.956 6.21e-06 ***
goog$Month02         0.001253   0.031531   0.040   0.9684
goog$Month03        -0.023391   0.032359  -0.723   0.4726
goog$Month04        -0.048513   0.032311  -1.501   0.1385
goog$Month05         0.002568   0.032431   0.079   0.9371
goog$Month06        -0.014763   0.032375  -0.456   0.6500
goog$Month07         0.074377   0.032477   2.290   0.0255 *
goog$Month08        -0.011228   0.032519  -0.345   0.7311
goog$Month09         0.030690   0.032339   0.949   0.3464
goog$Month10         0.048443   0.033136   1.462   0.1490
goog$Month11        -0.012551   0.032330  -0.388   0.6992
goog$Month12         0.025718   0.032315   0.796   0.4293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05596 on 60 degrees of freedom
Multiple R-squared:  0.5187,    Adjusted R-squared:  0.4225
F-statistic: 5.389 on 12 and 60 DF,  p-value: 4.221e-06
```

THESE STATISTICS ARE CALCULATED ASSUMING THAT THE RESIDUALS ARE NORMALLY DISTRIBUTED

ONE OF THE DIAGNOSTIC PLOTS IS A NORMAL Q-Q PLOT

ASSUMPTION 1:
THE RESIDUALS ARE NORMALLY DISTRIBUTED

ONE OF THE DIAGNOSTIC PLOTS IS **A NORMAL Q-Q PLOT**



ASSUMPTION 1:

THE RESIDUALS ARE NORMALLY DISTRIBUTED

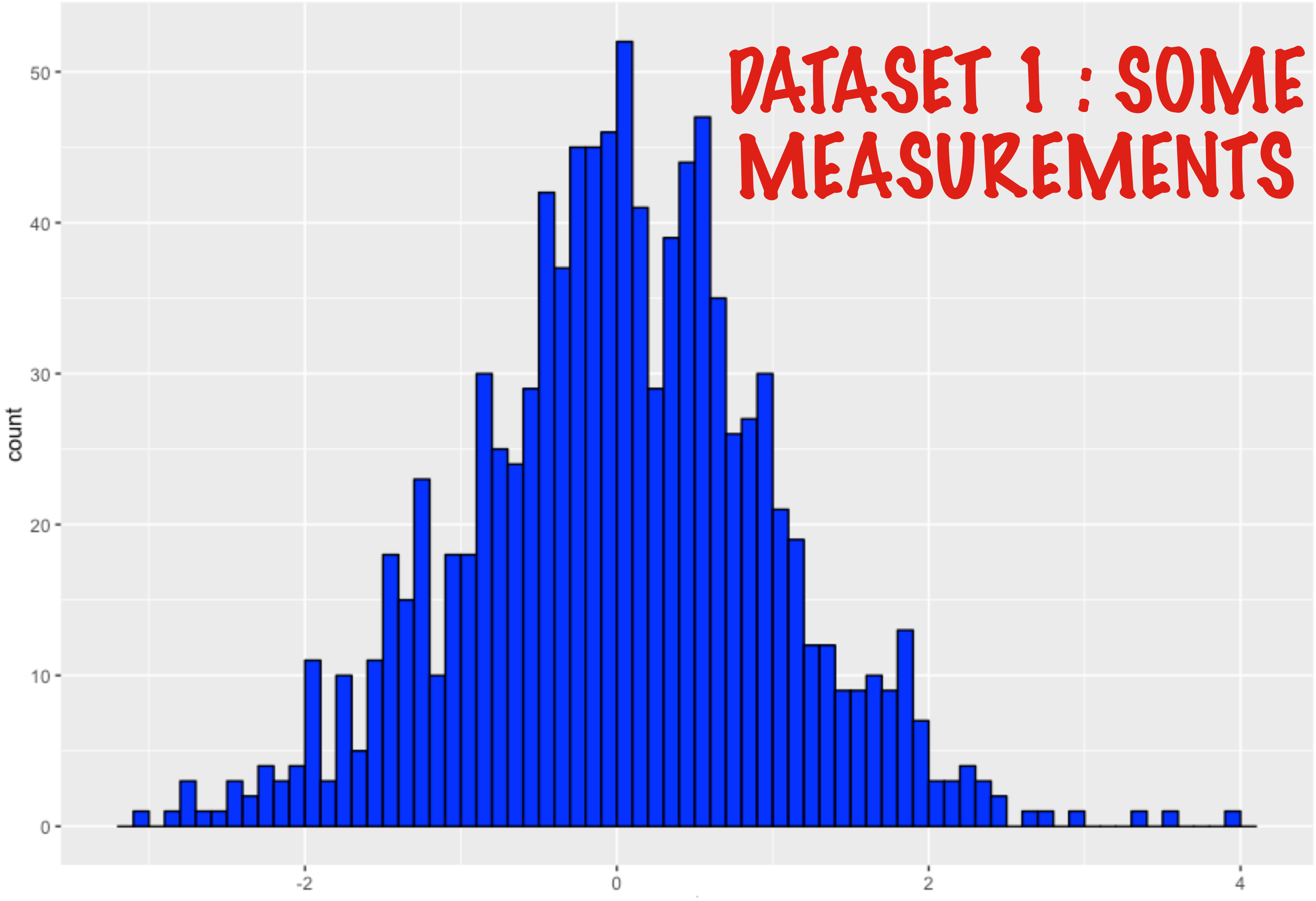
A Q-Q PLOT

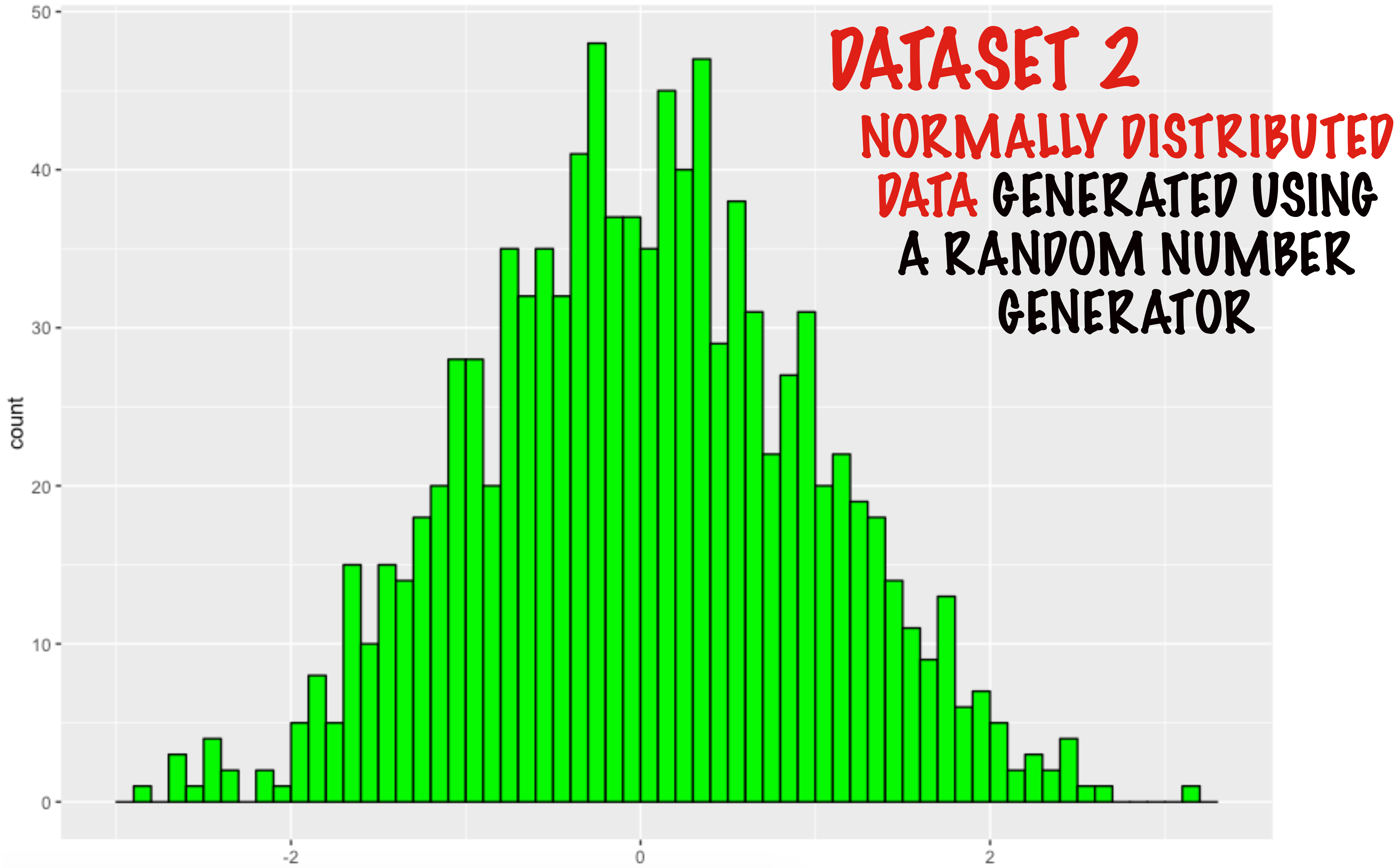
(QUANTILE-QUANTILE PLOT)

IS A VISUAL WAY OF CHECKING WHETHER SOME DATA FITS A PARTICULAR DISTRIBUTION

A Q-Q PLOT

IS A ROUNDABOUT WAY TO COMPARE THE DISTRIBUTIONS (HISTOGRAMS) OF 2 DATASETS





DO THESE 2 DATASETS HAVE THE SAME PROBABILITY DISTRIBUTION?

A Q-Q PLOT

A Q-Q PLOT COMPARES QUANTILES OF THE DATASETS

QUANTILES ARE POINTS THAT DIVIDE THE DATA (ONCE IT'S SORTED) INTO EQUAL SIZED GROUPS

QUARTILES DIVIDE THE DATA INTO 4 EQUAL GROUPS

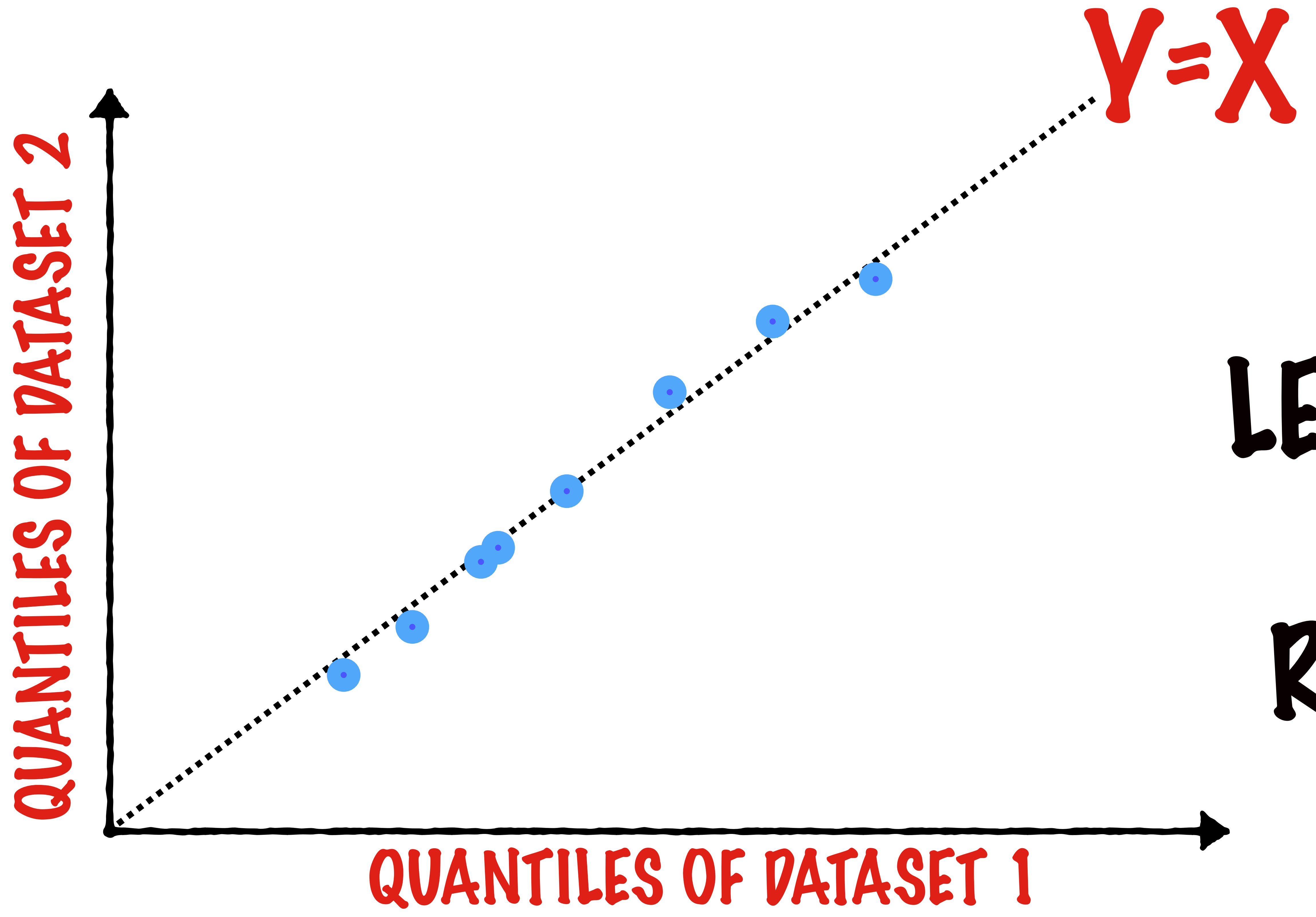
PERCENTILES DIVIDE THE DATA INTO 100 EQUAL GROUPS

THE IDEA IS IF THE QUANTILES OF THE 2 DATASETS ARE EQUAL, THEN THEY ARE FROM THE SAME DISTRIBUTION

A Q-Q PLOT

A Q-Q PLOT COMPARES QUANTILES OF THE DATASETS

IF THE QUANTILES ARE EQUAL THEY WILL LIE ON THE LINE Y=X



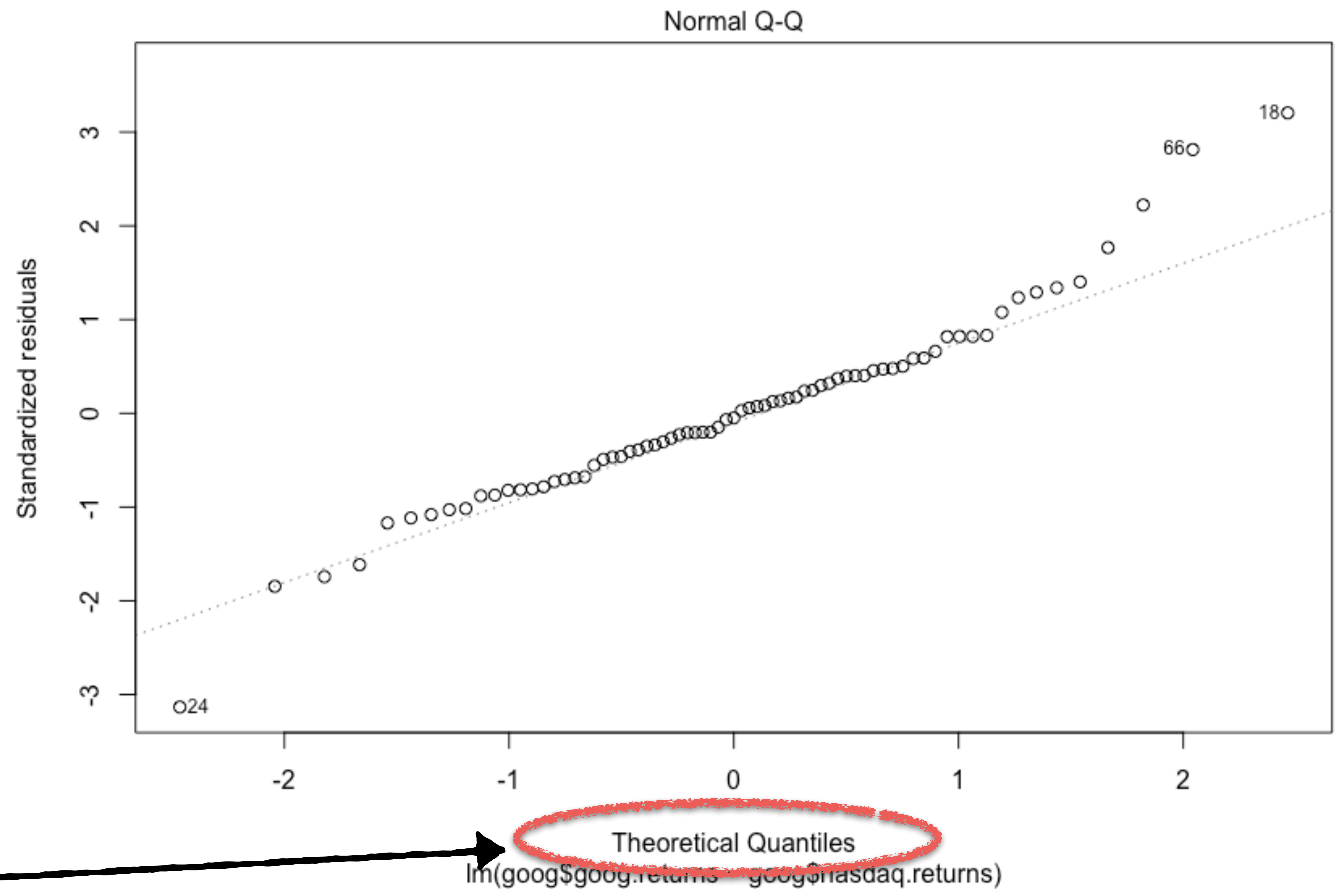
LET'S GO BACK TO LINEAR REGRESSION

ASSUMPTION 1:

THE RESIDUALS ARE NORMALLY DISTRIBUTED

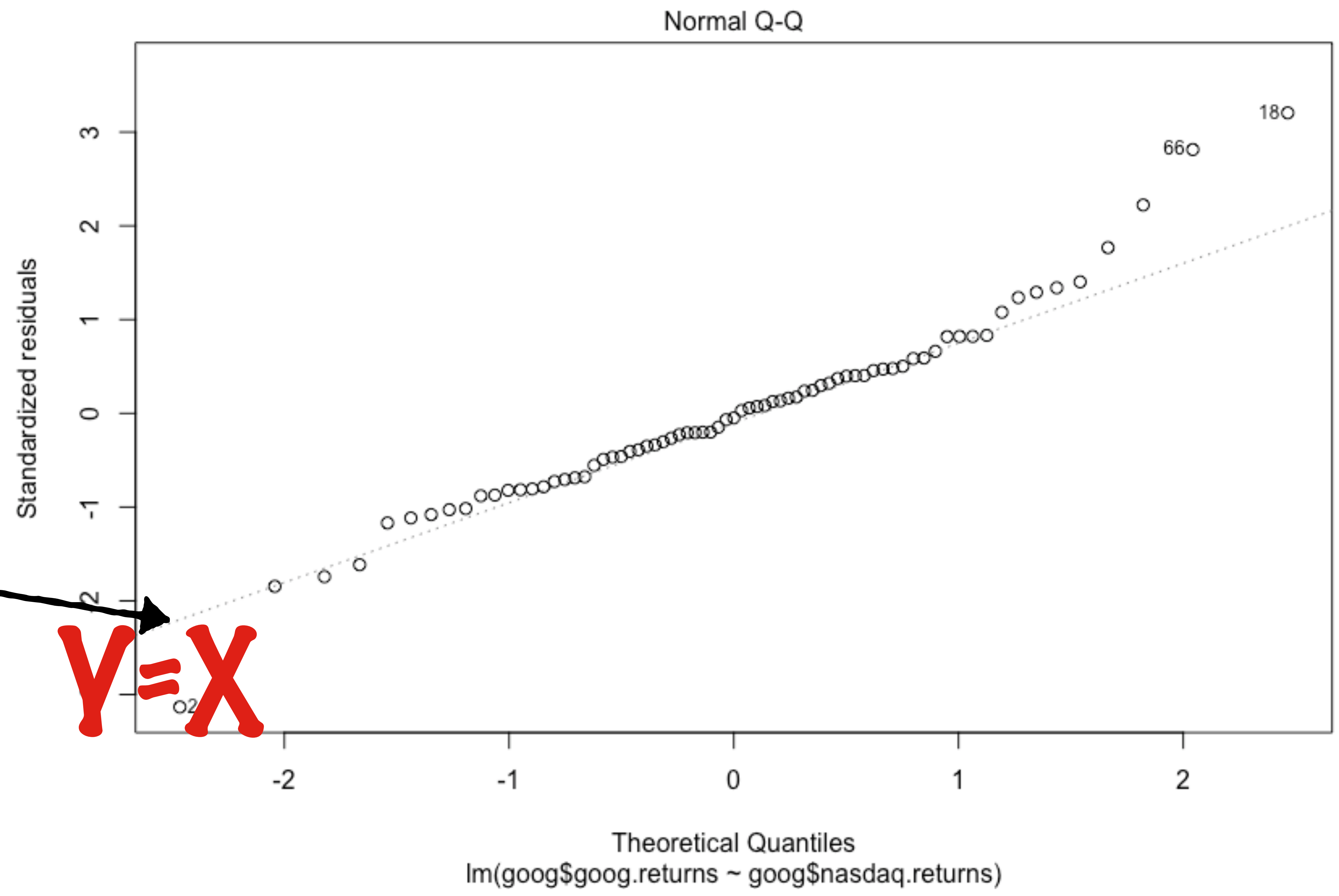
ONE OF THE DIAGNOSTIC PLOTS IS A NORMAL Q-Q PLOT

IT PLOTS THE QUANTILES OF THE RESIDUALS AGAINST QUANTILES FROM A NORMAL DISTRIBUTION



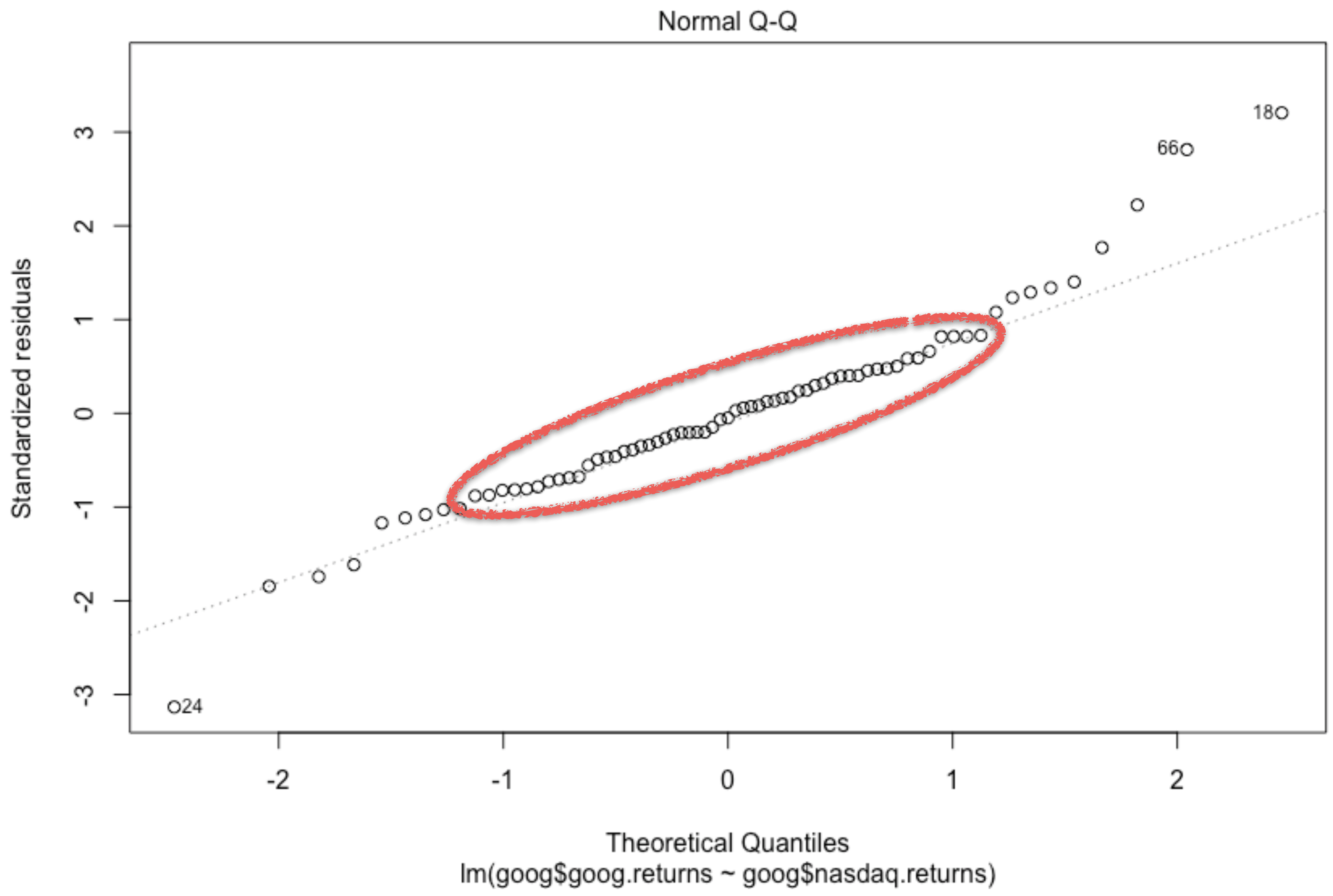
ASSUMPTION 1:
THE RESIDUALS ARE NORMALLY DISTRIBUTED

FOR OUR ASSUMPTION TO BE TRUE ALL THE POINTS NEED TO LIE ALONG THE DOTTED LINE



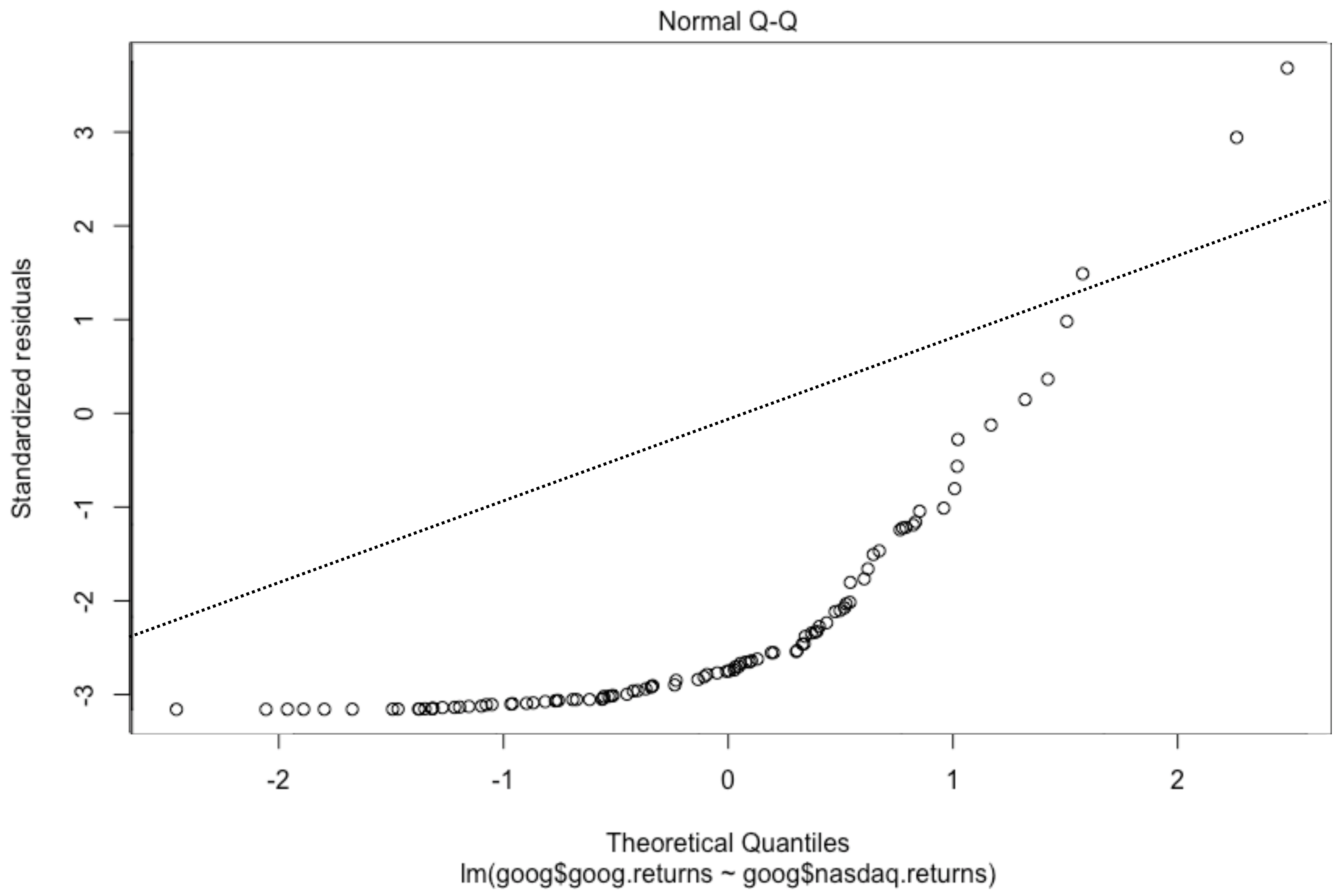
ASSUMPTION 1:
THE RESIDUALS ARE NORMALLY DISTRIBUTED

MOST OF THE POINTS IN THIS GRAPH SUPPORT OUR ASSUMPTION



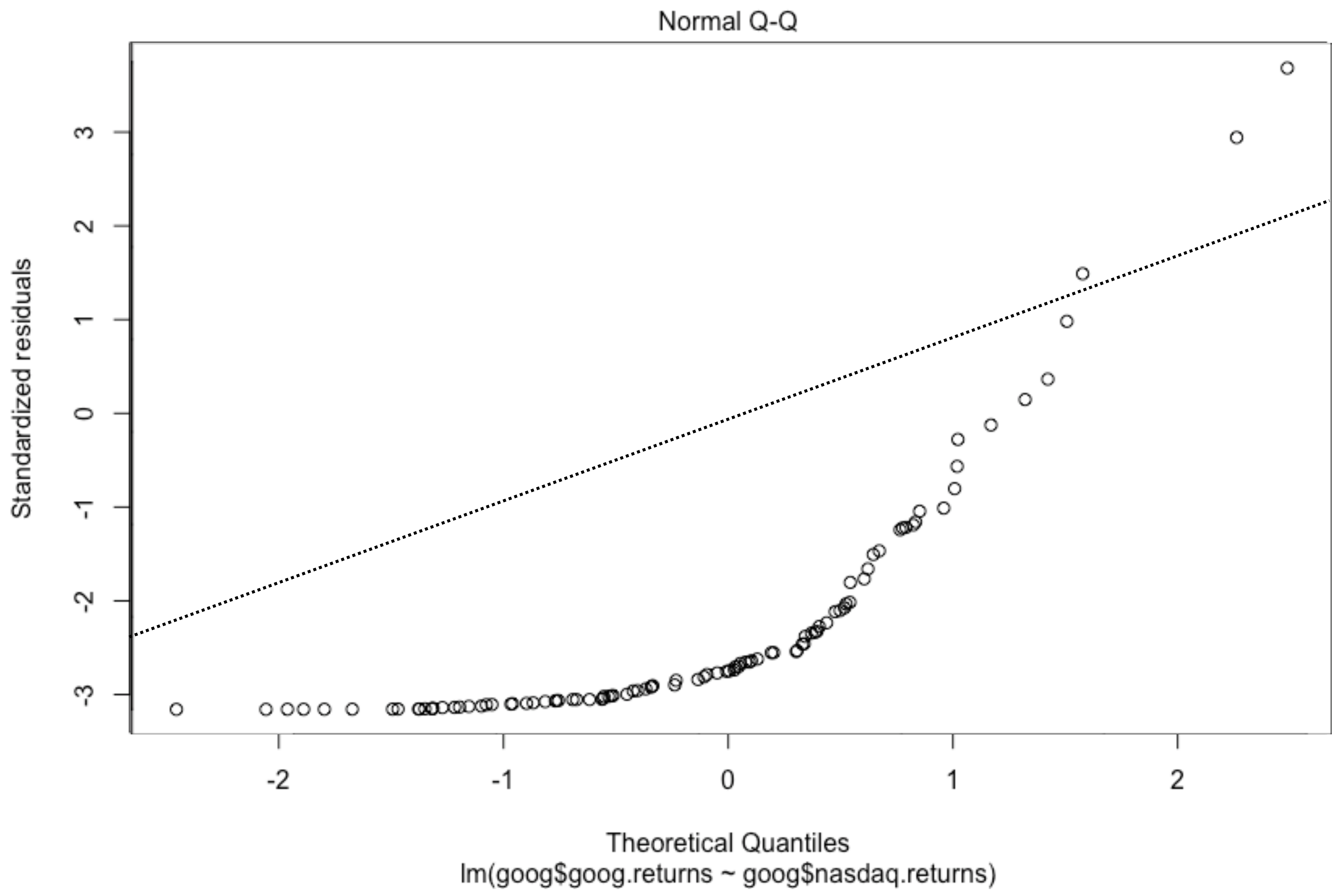
ASSUMPTION 1:
THE RESIDUALS ARE NORMALLY DISTRIBUTED

HERE IS AN EXAMPLE THAT DOES NOT SUPPORT OUR ASSUMPTION



ASSUMPTION 1:
THE RESIDUALS ARE NORMALLY DISTRIBUTED

IF YOU SEE A RESIDUAL Q-Q PLOT LIKE THIS, THEN YOUR LINEAR MODEL WOULD NOT BE VALID I.E. NOT A GOOD REPRESENTATION OF THE DATA



ASSUMPTION 2:

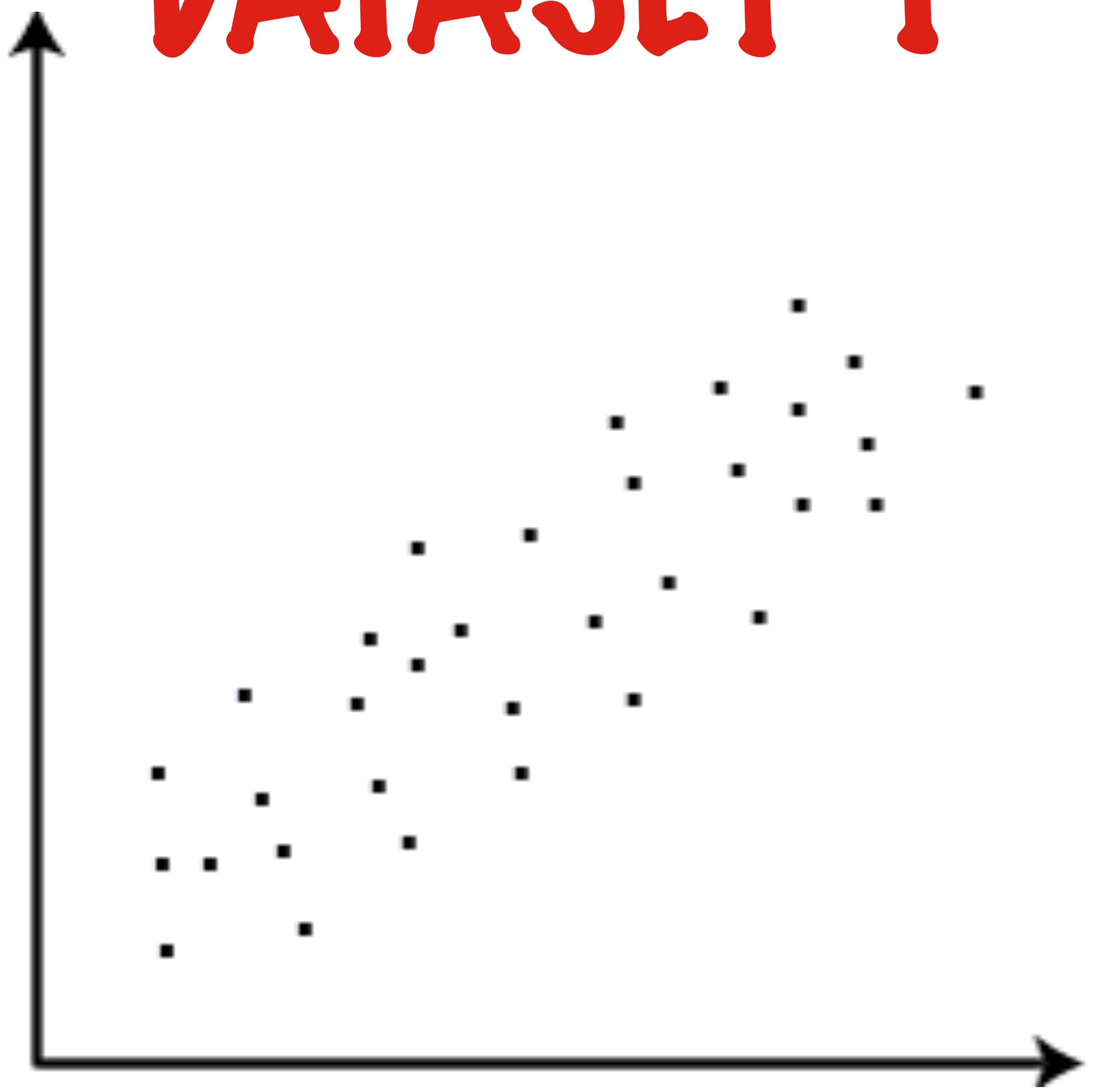
THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE WITH RESPECT TO THE FITTED LINE

ASSUMPTION 2:

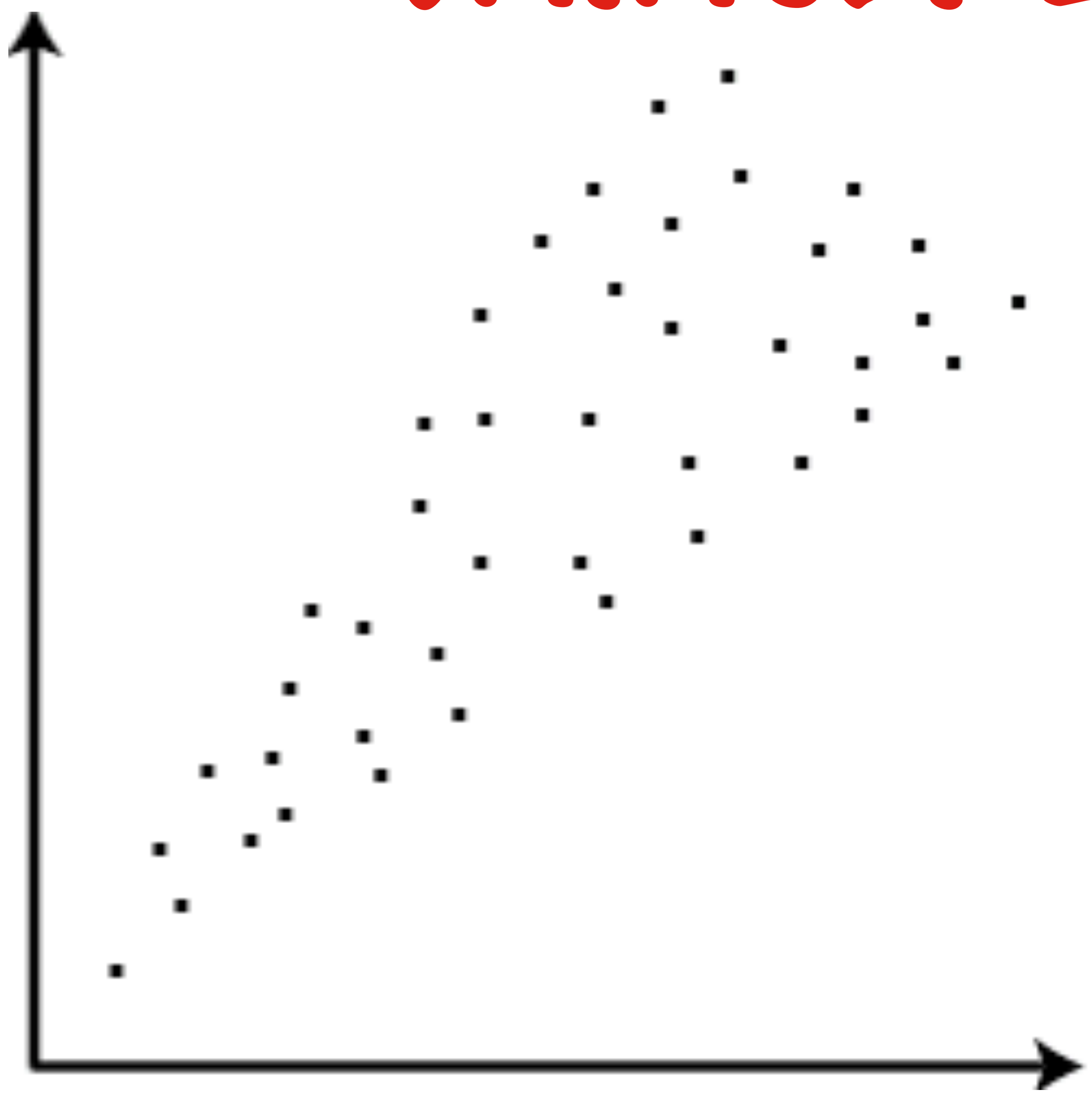
THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

HERE ARE 2 DATASETS

DATASET 1



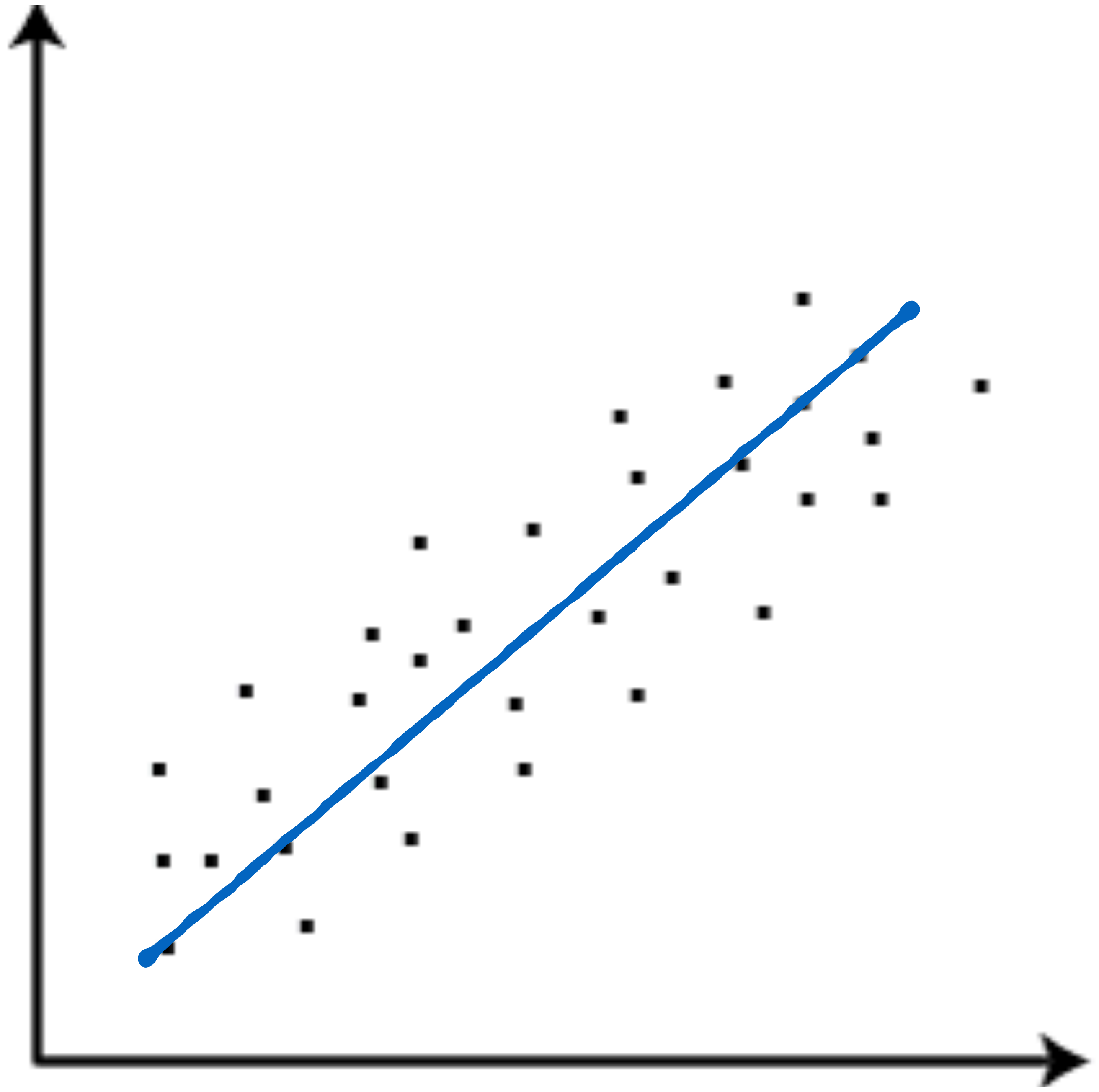
DATASET 2



ASSUMPTION 2:

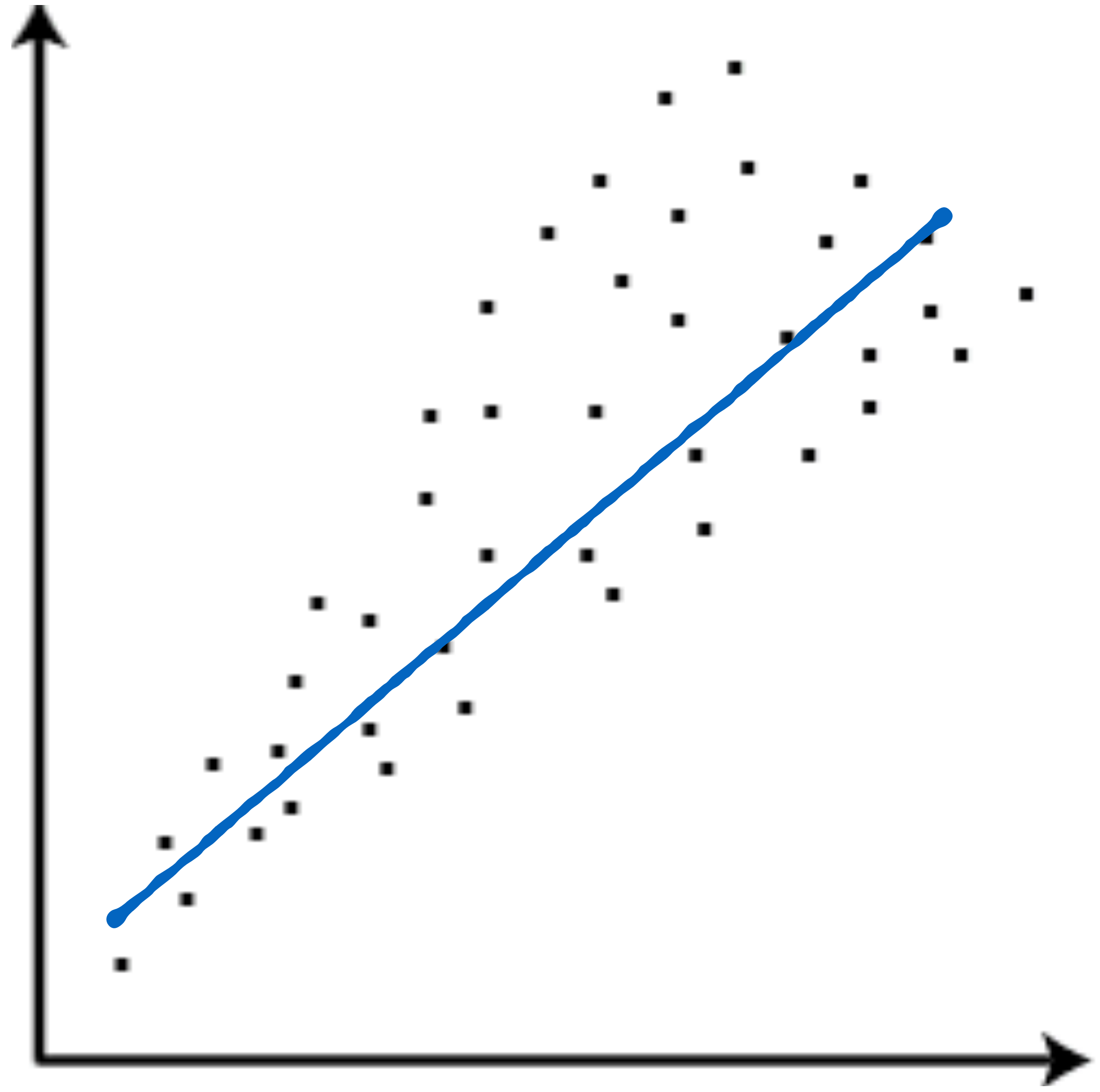
THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

DATASET 1



$$\hat{Y} = \beta_0 + \beta_1 X_1$$

DATASET 2



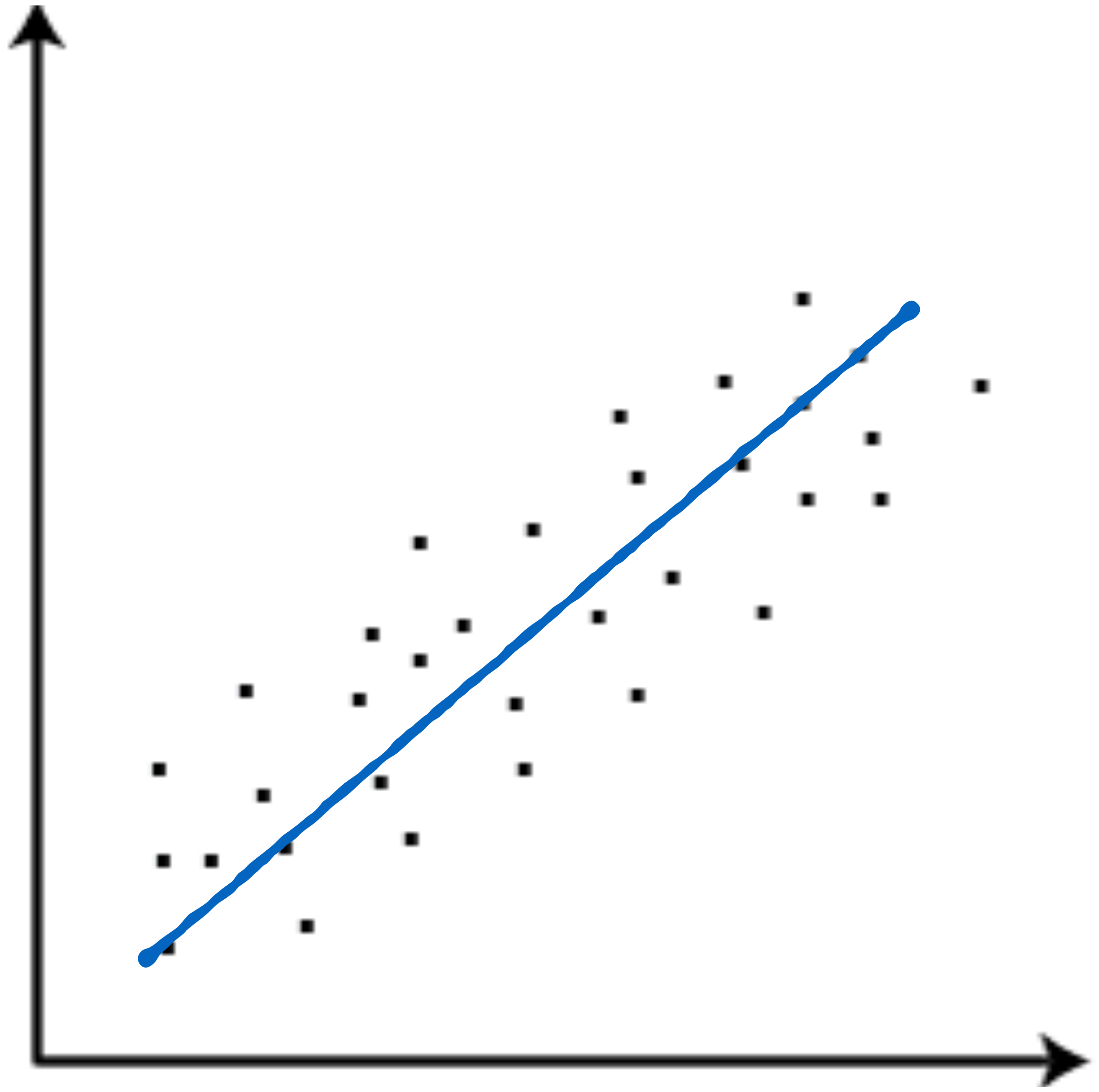
$$\hat{Y} = \beta_0 + \beta_1 X_1$$

WHEN YOU PERFORM LINEAR REGRESSION, BOTH DATASETS GIVE YOU THE SAME LINE

ASSUMPTION 2:

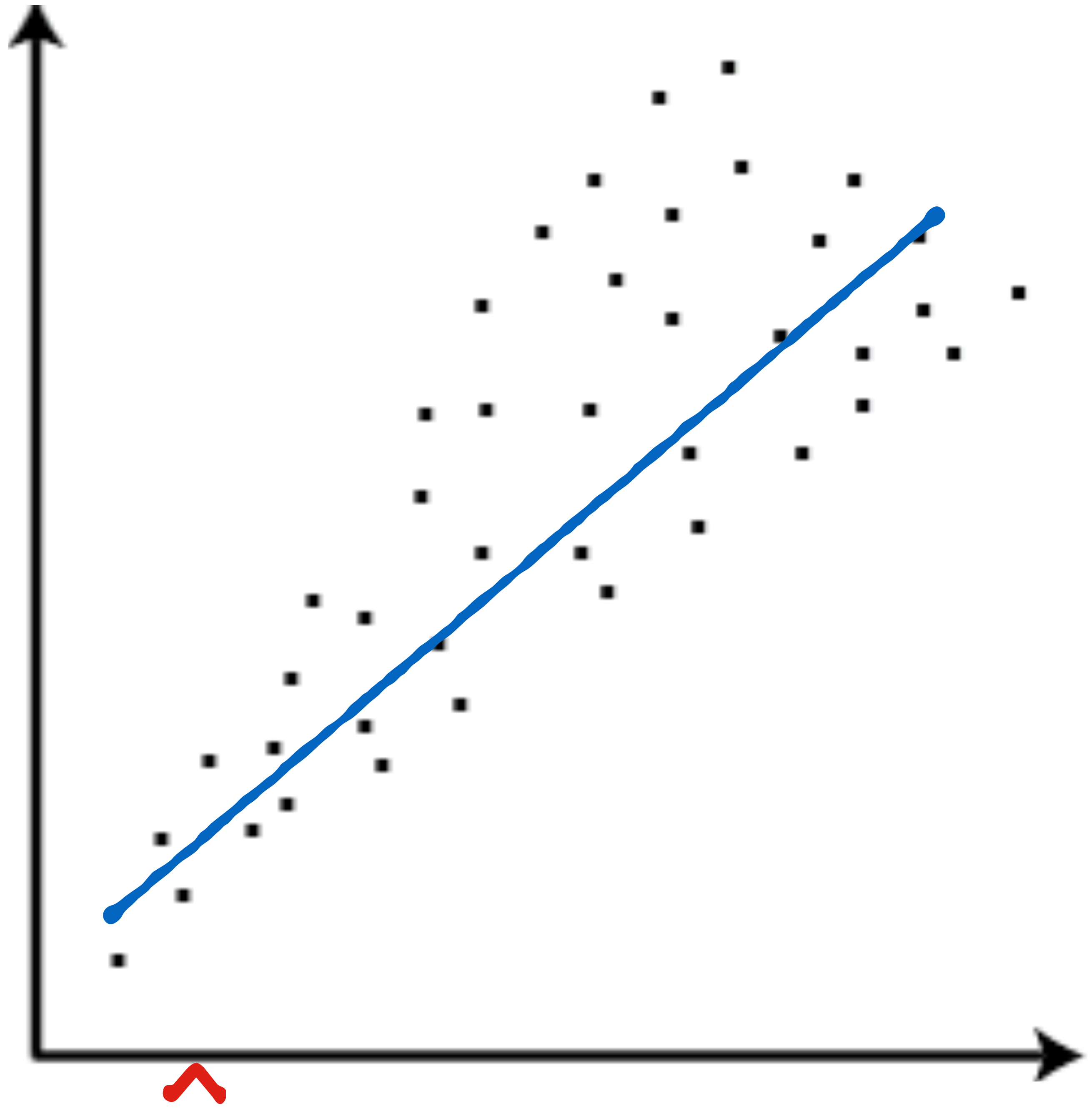
THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

DATASET 1



$$\hat{Y} = \beta_0 + \beta_1 X_1$$

DATASET 2



$$\hat{Y} = \beta_0 + \beta_1 X_1$$

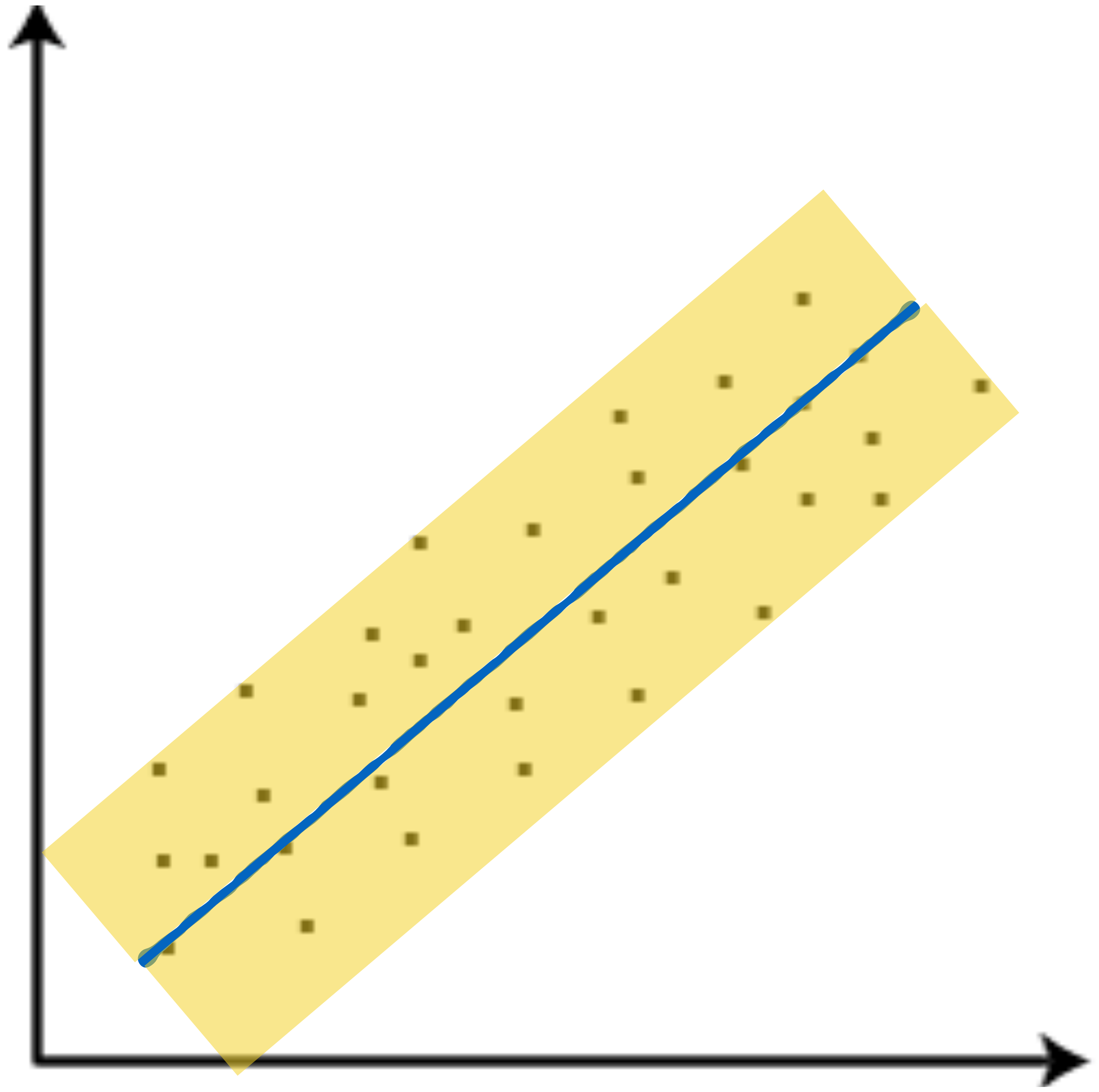
THE LINE IS A BETTER REPRESENTATION FOR DATASET 1

WHY?

ASSUMPTION 2:

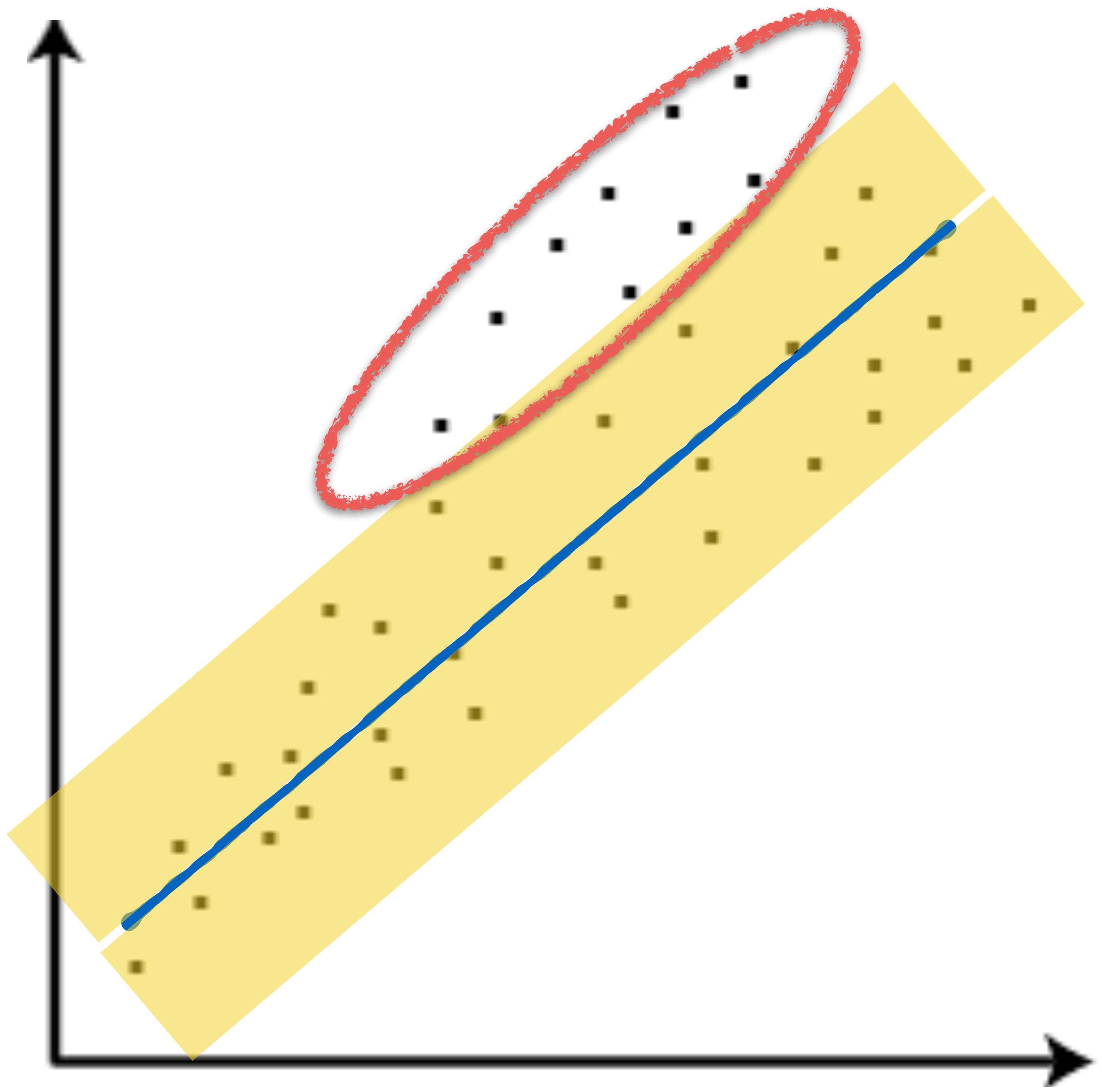
THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

DATASET 1



ALL POINTS
FALL WITHIN
THIS INTERVAL

DATASET 2

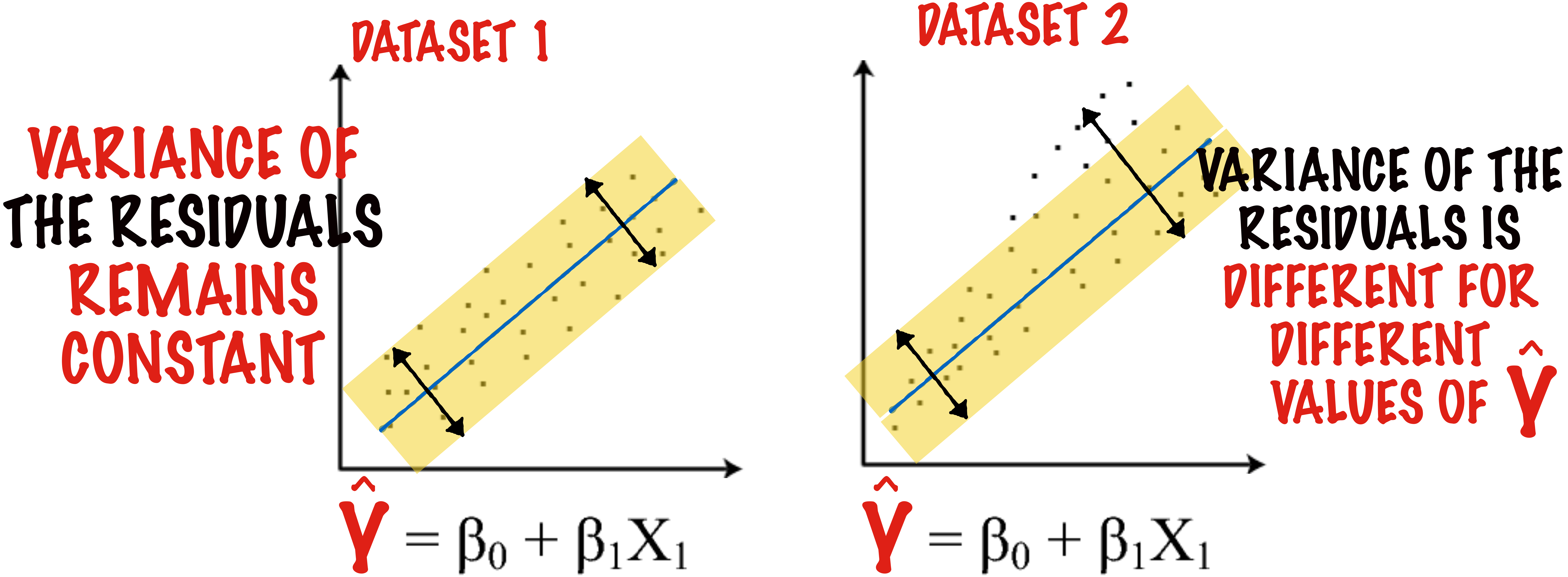


SOME POINTS
FALL OUTSIDE
THIS INTERVAL

LET'S DRAW A CONSTANT INTERVAL AROUND THE FITTED LINE

ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

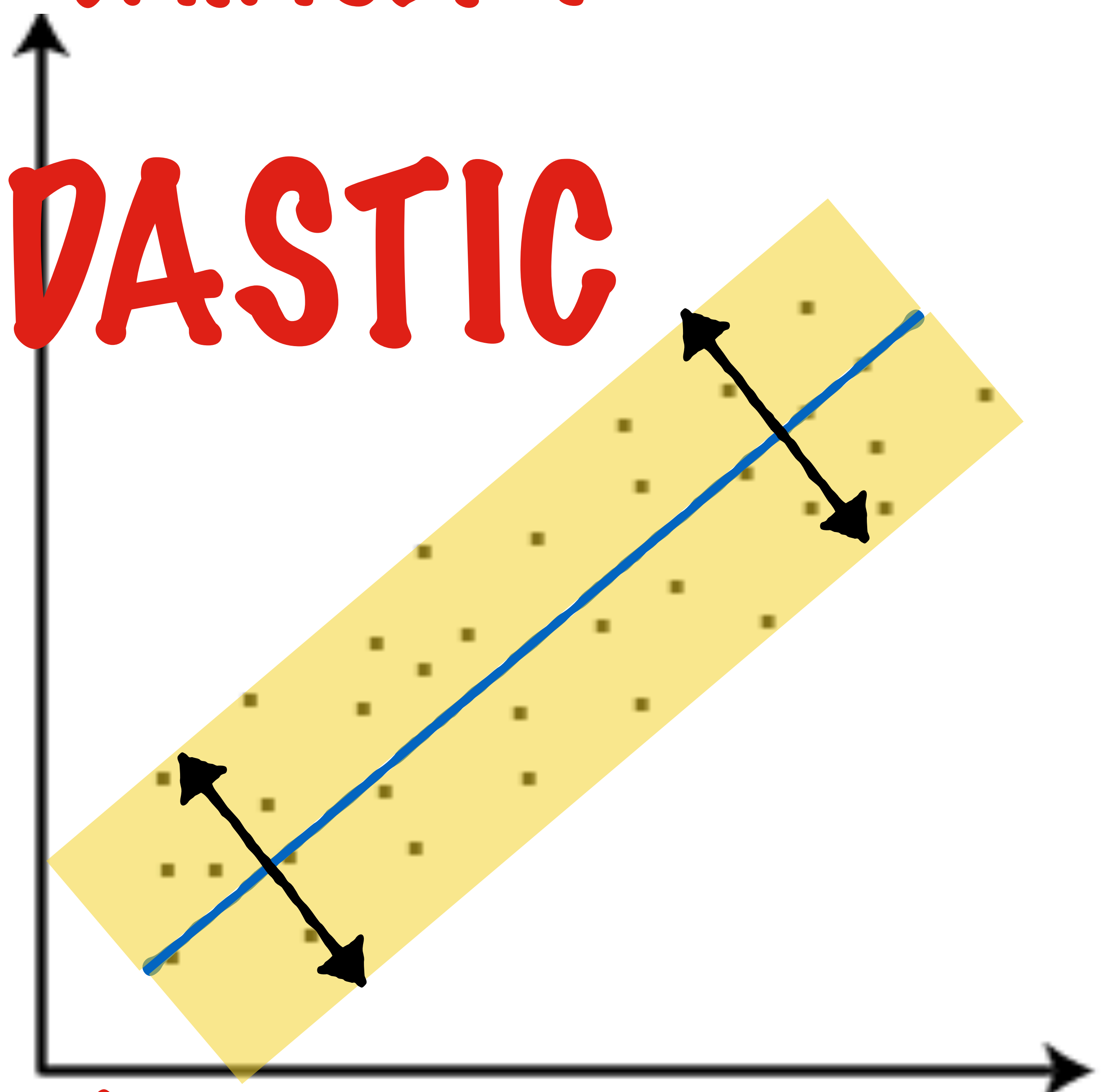


ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

DATASET 1

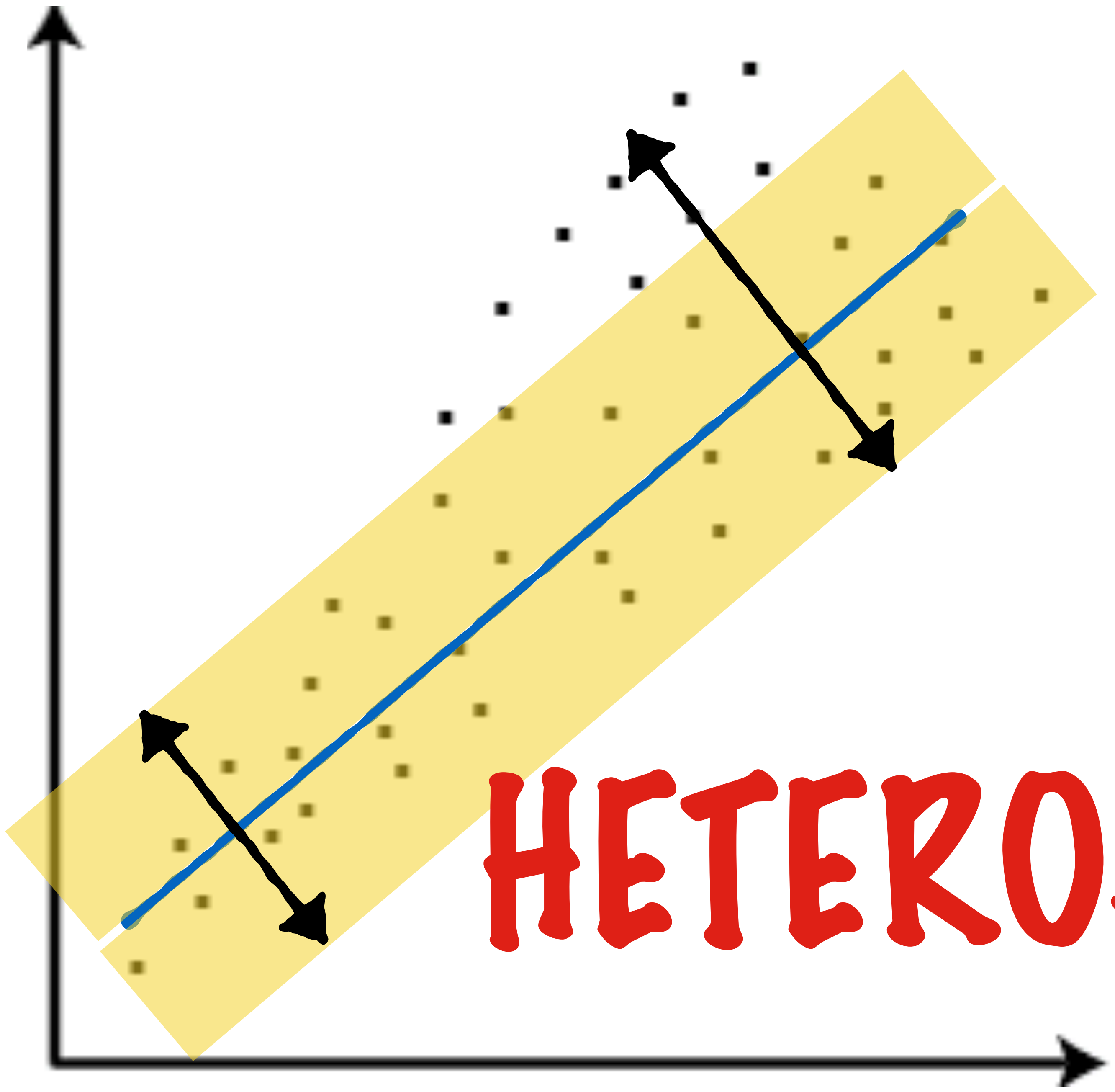
HOMOSCEDASTIC



$$\hat{Y} = \beta_0 + \beta_1 X_1$$

DATASET 2

HETEROSCEDASTIC



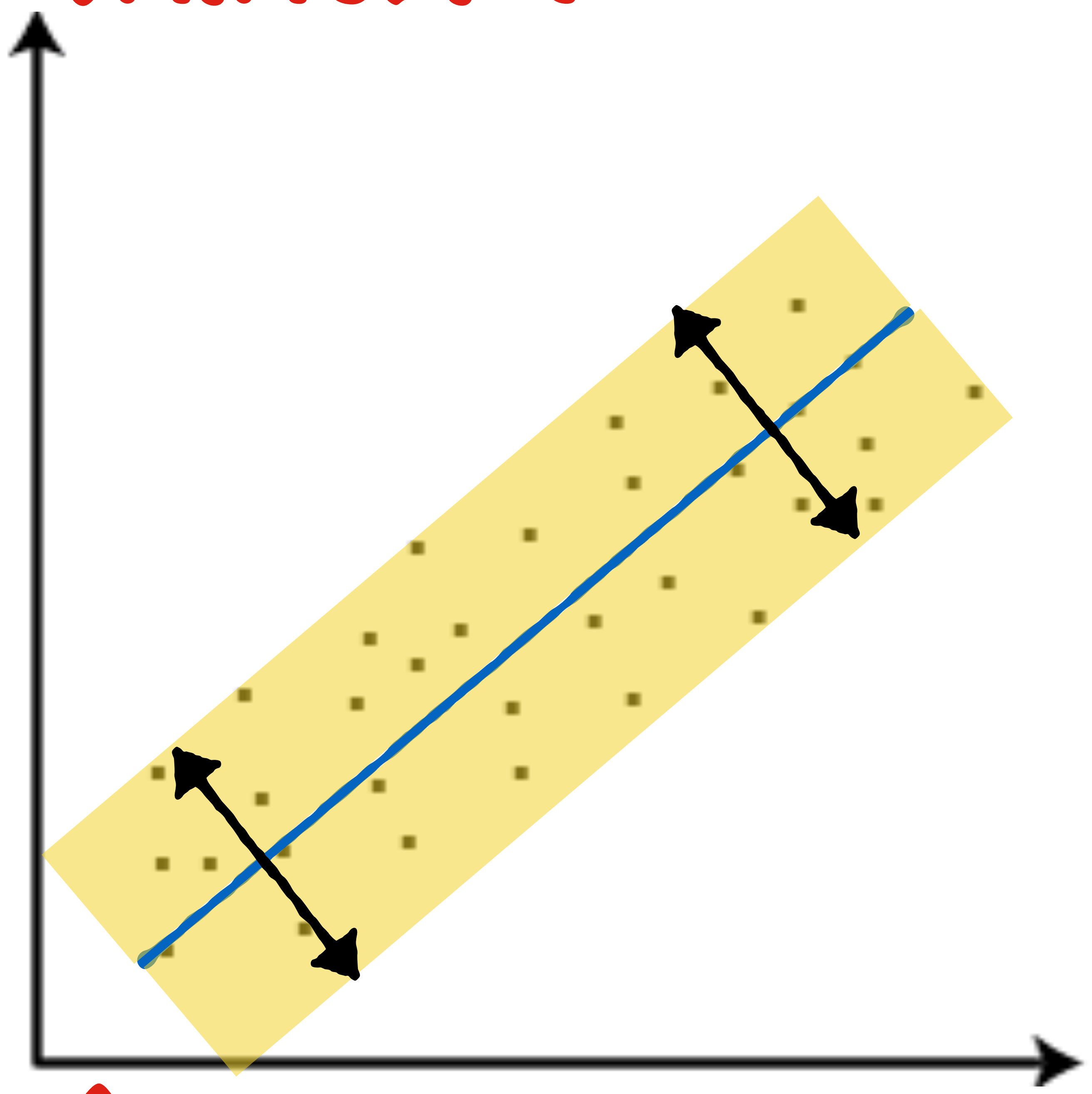
$$\hat{Y} = \beta_0 + \beta_1 X_1$$

ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

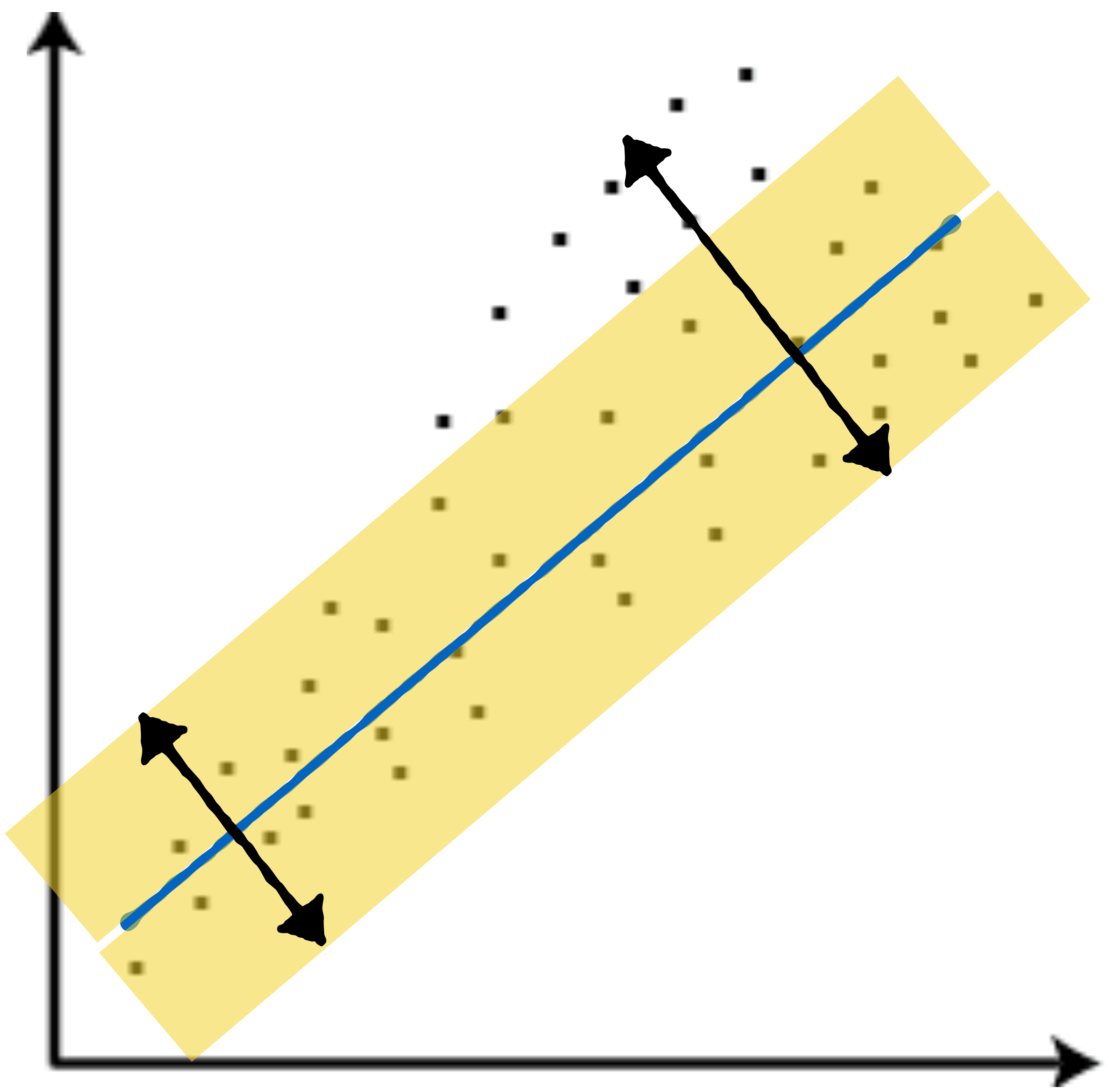
THERE ARE 2 DIAGNOSTIC PLOTS THAT HELP US CHECK IF OUR DATA IS MORE LIKE DATASET1 OR DATASET2

DATASET 1



$\hat{Y} = \beta_0 + \beta_1 X_1$

DATASET 2

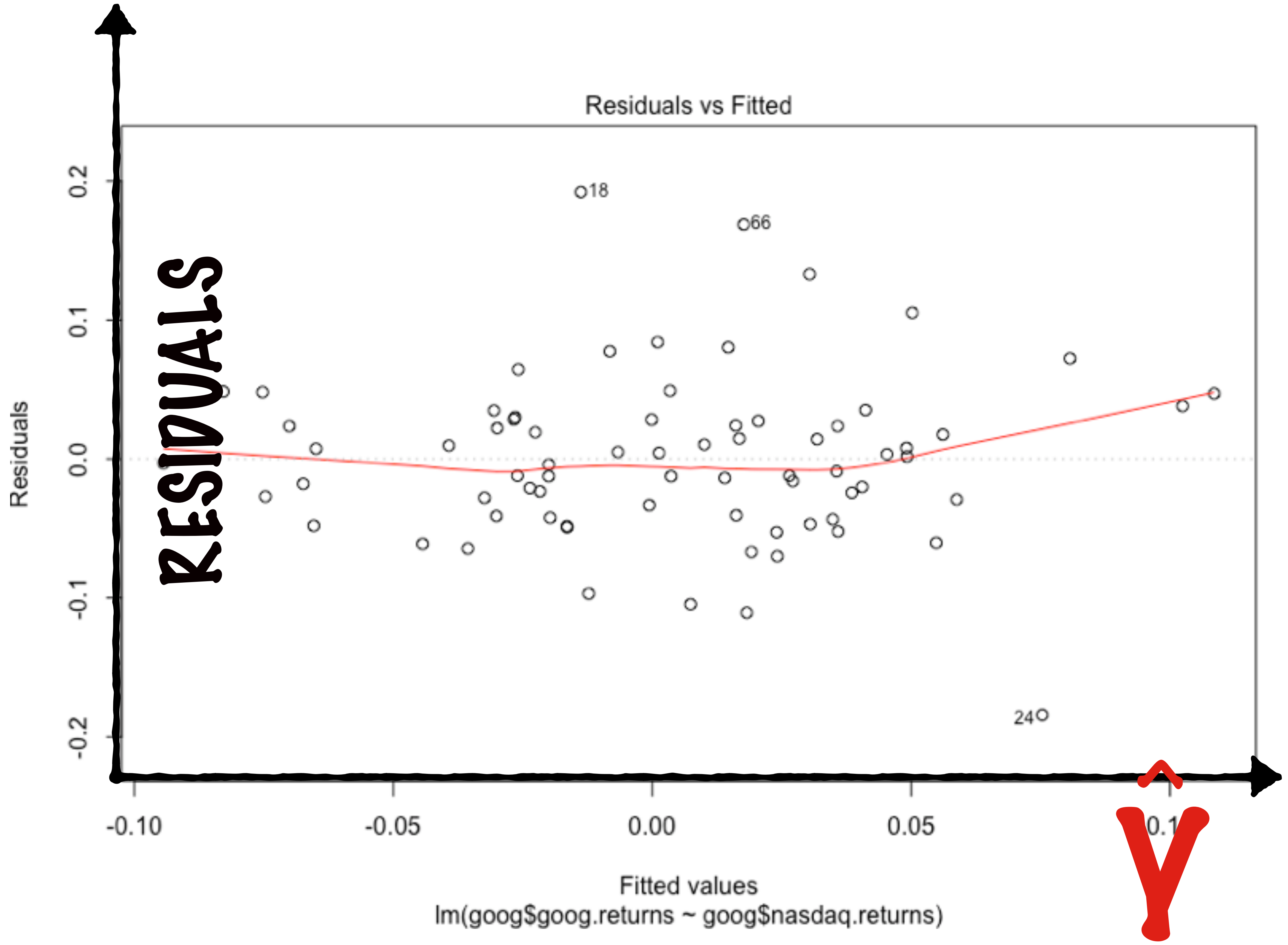


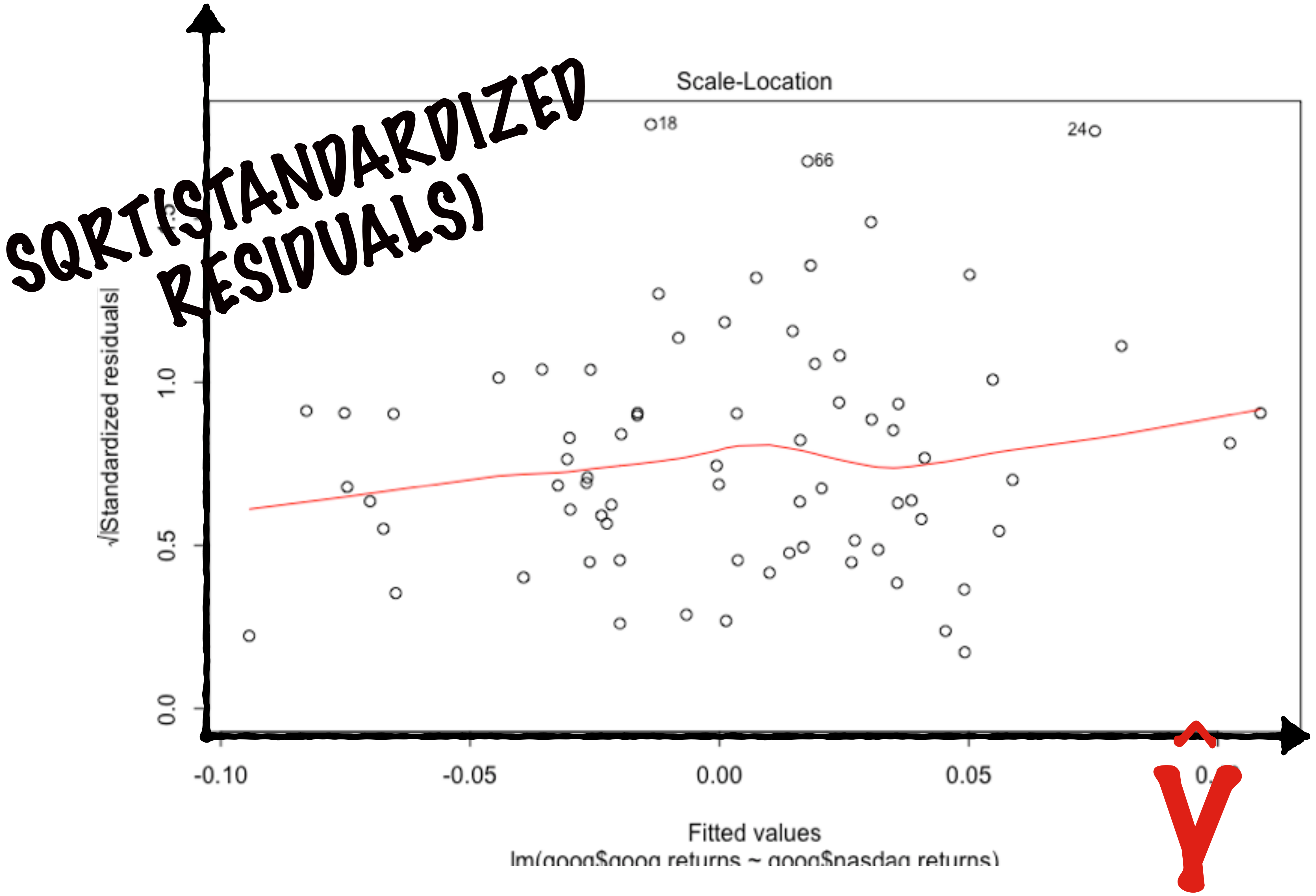
$\hat{Y} = \beta_0 + \beta_1 X_1$

ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

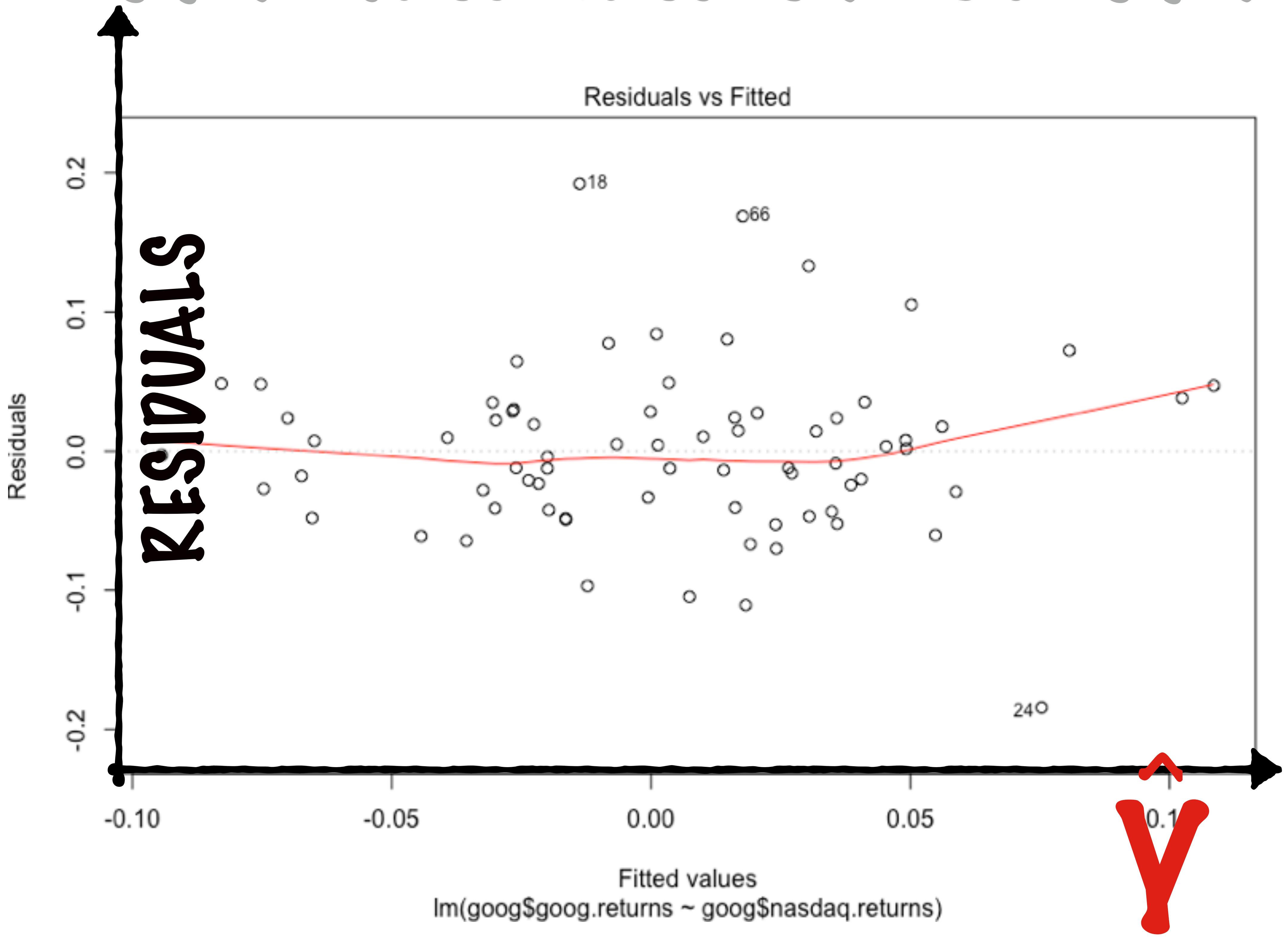
2 DIAGNOSTIC PLOTS

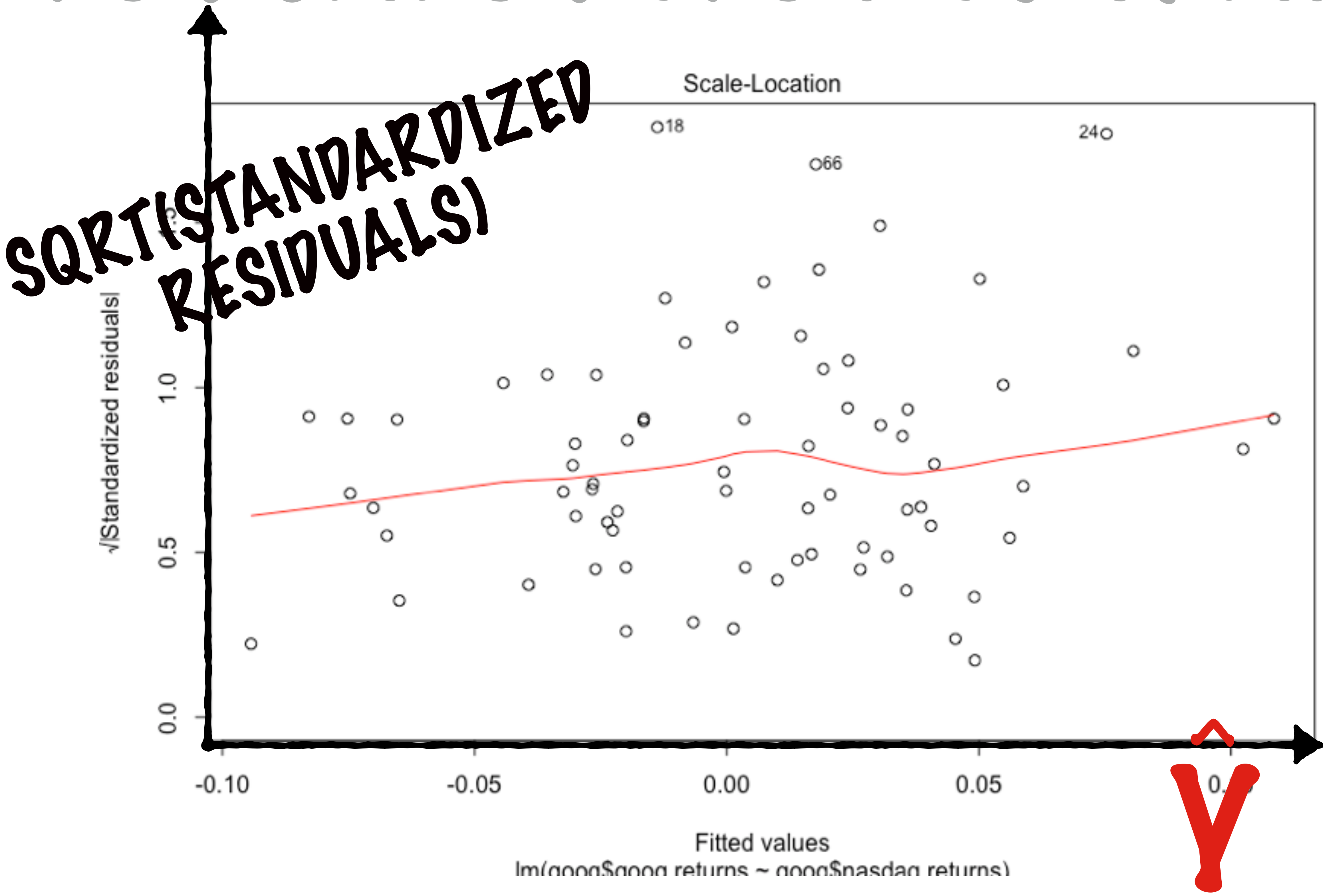




ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

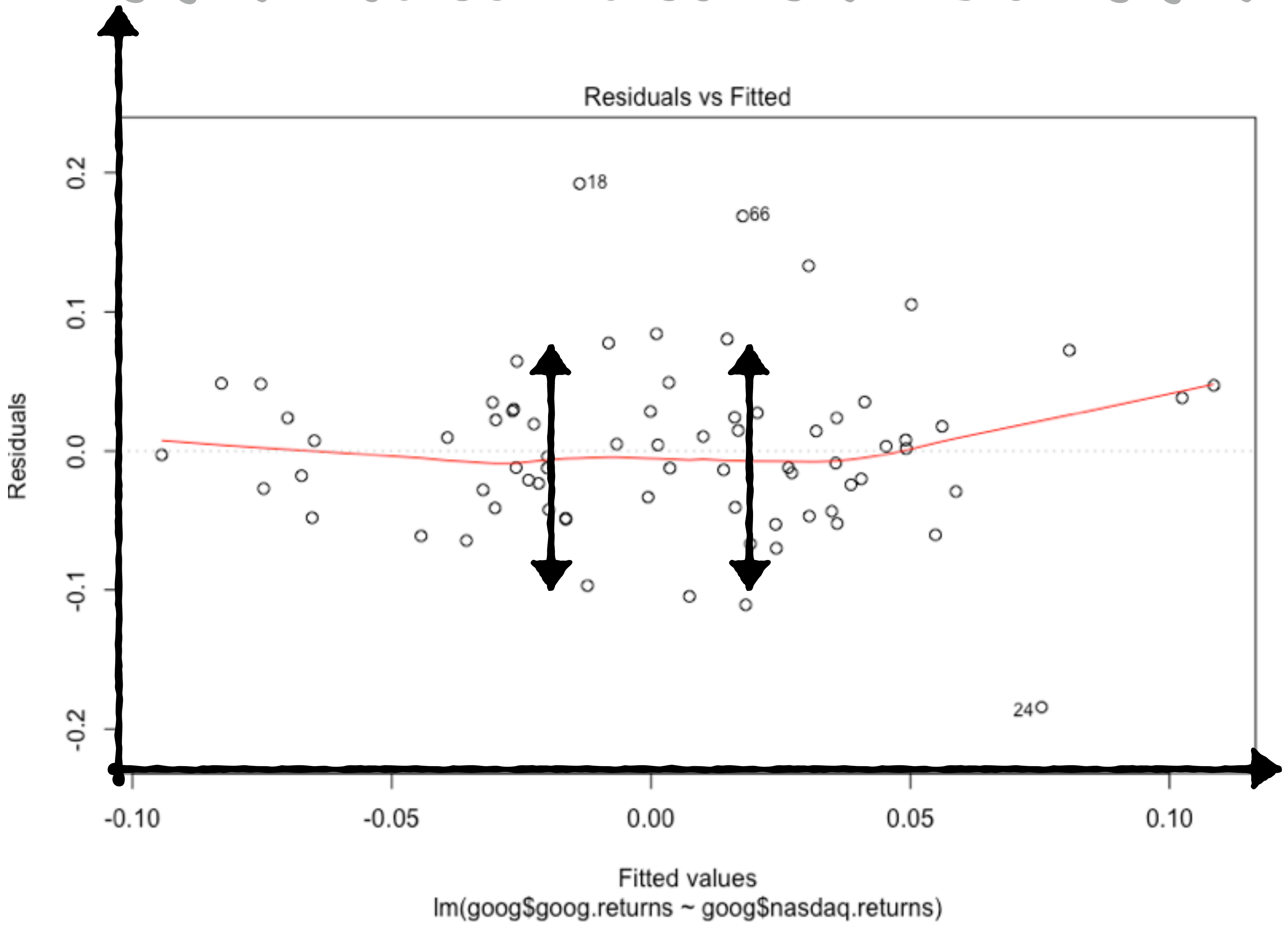


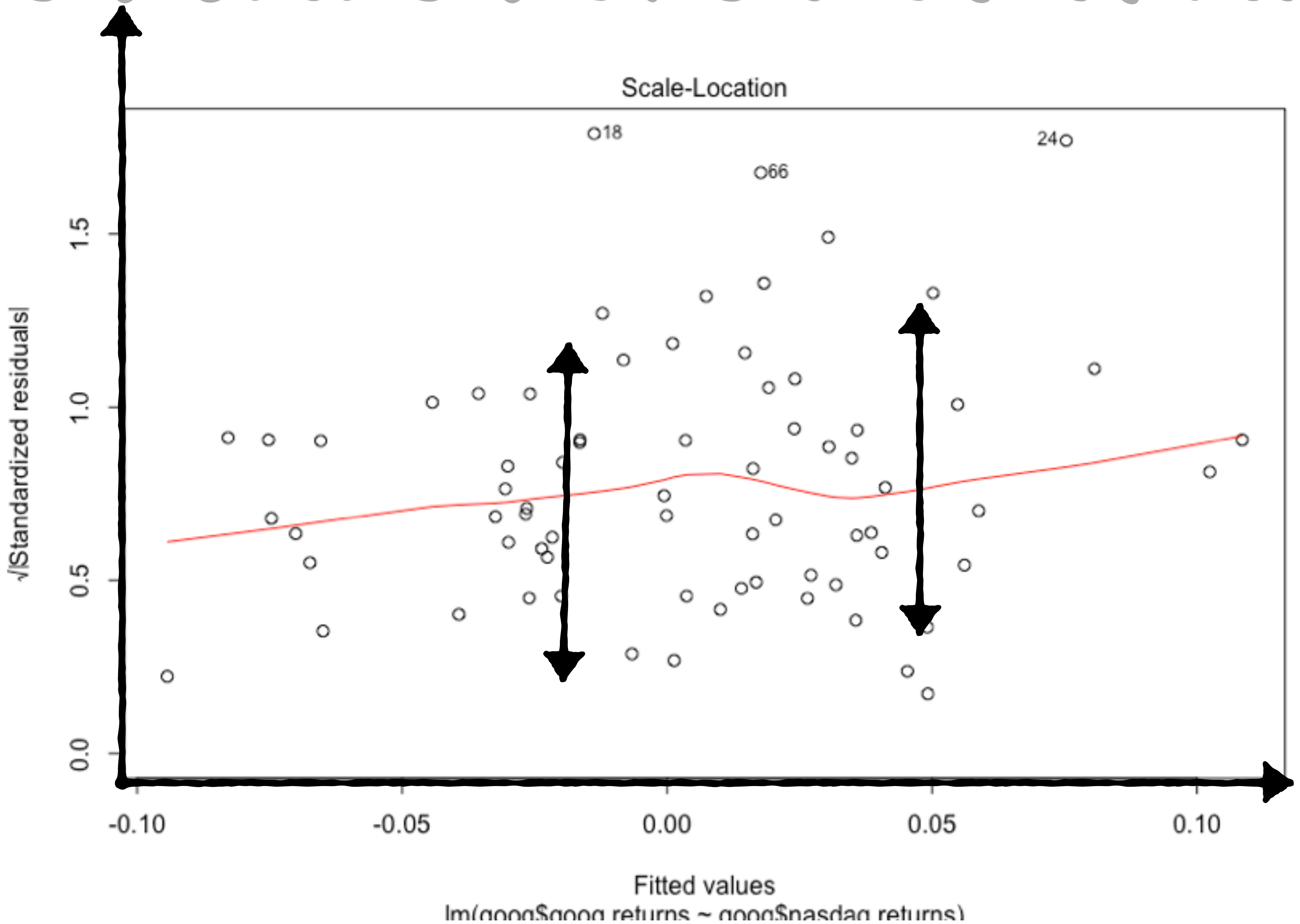


STANDARDIZED JUST MEANS THAT **RESIDUALS HAVE BEEN SCALED** TO FIT THE STANDARD NORMAL DISTRIBUTION (MEAN 0, SD 1)

ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

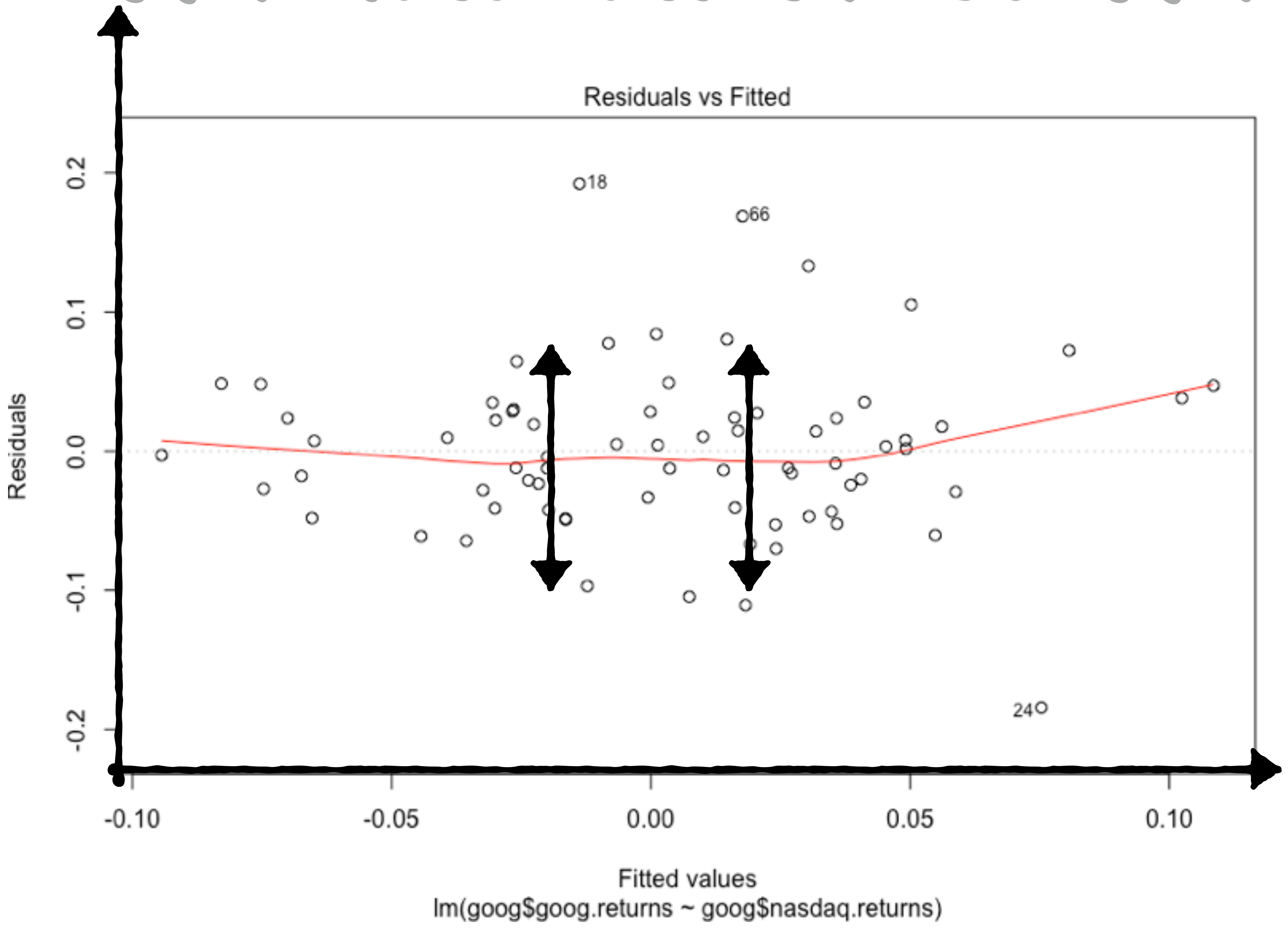


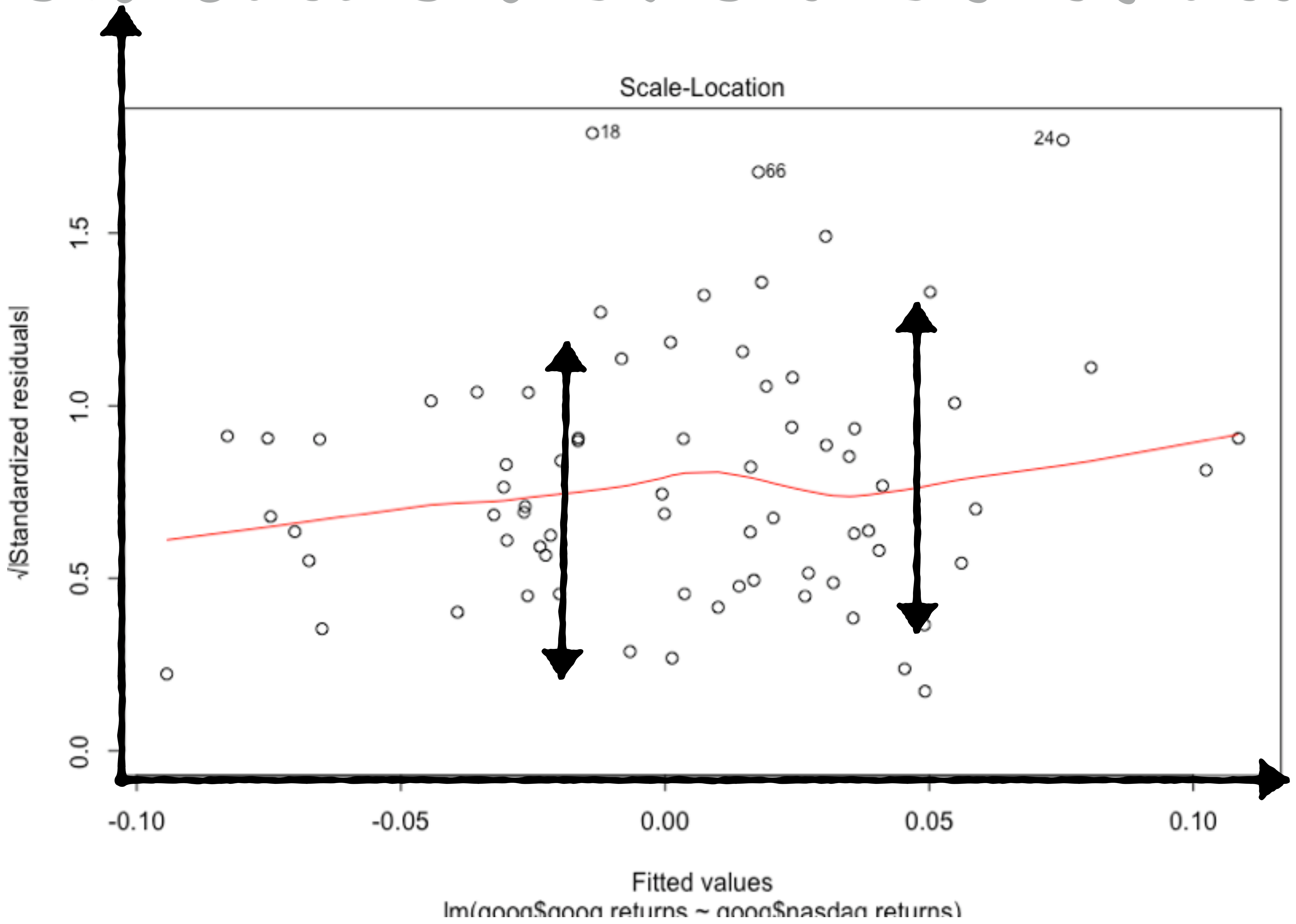


FOR VALIDATING OUR ASSUMPTION: JUST CHECK IF THE SPREAD REMAINS CONSTANT

ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE



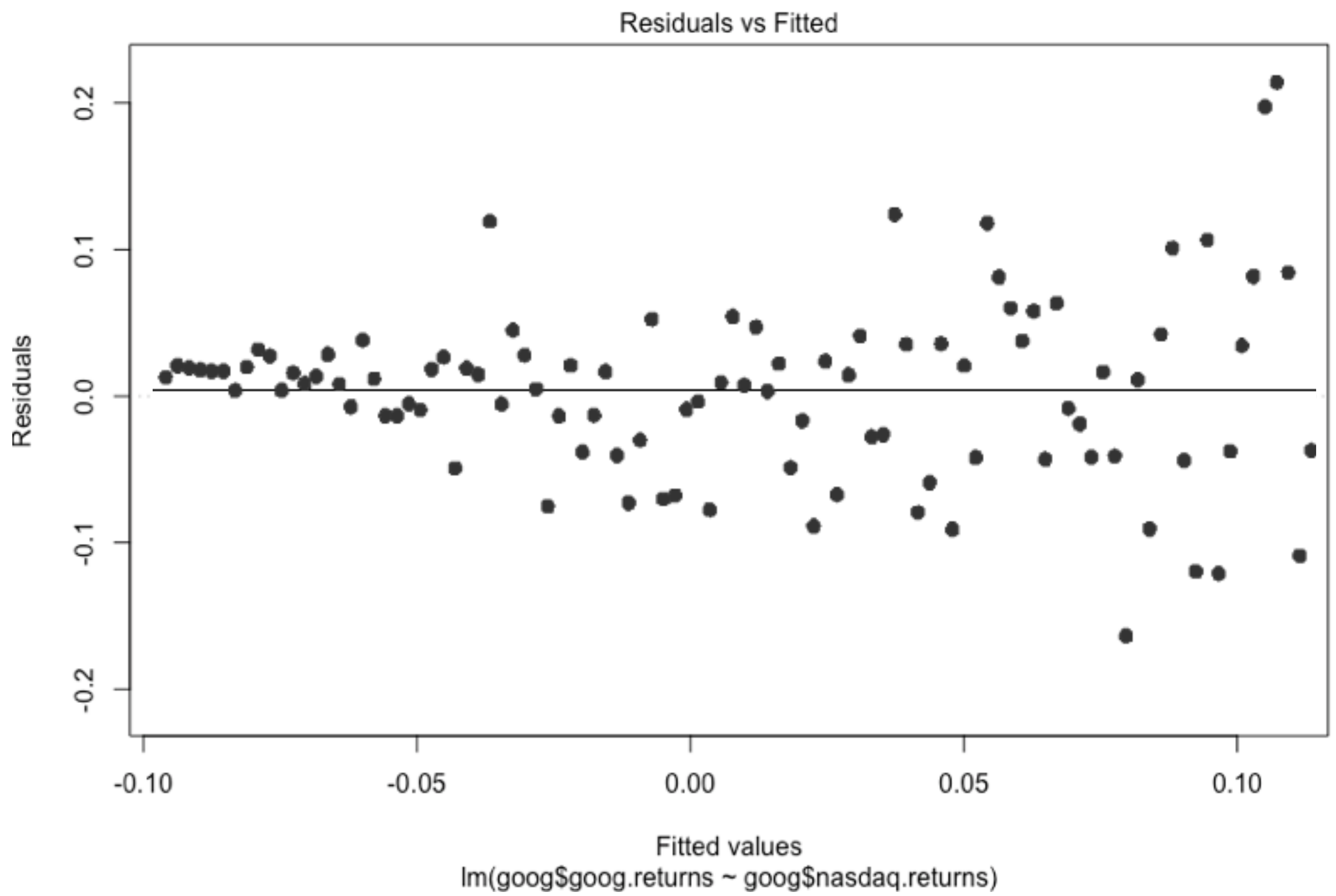


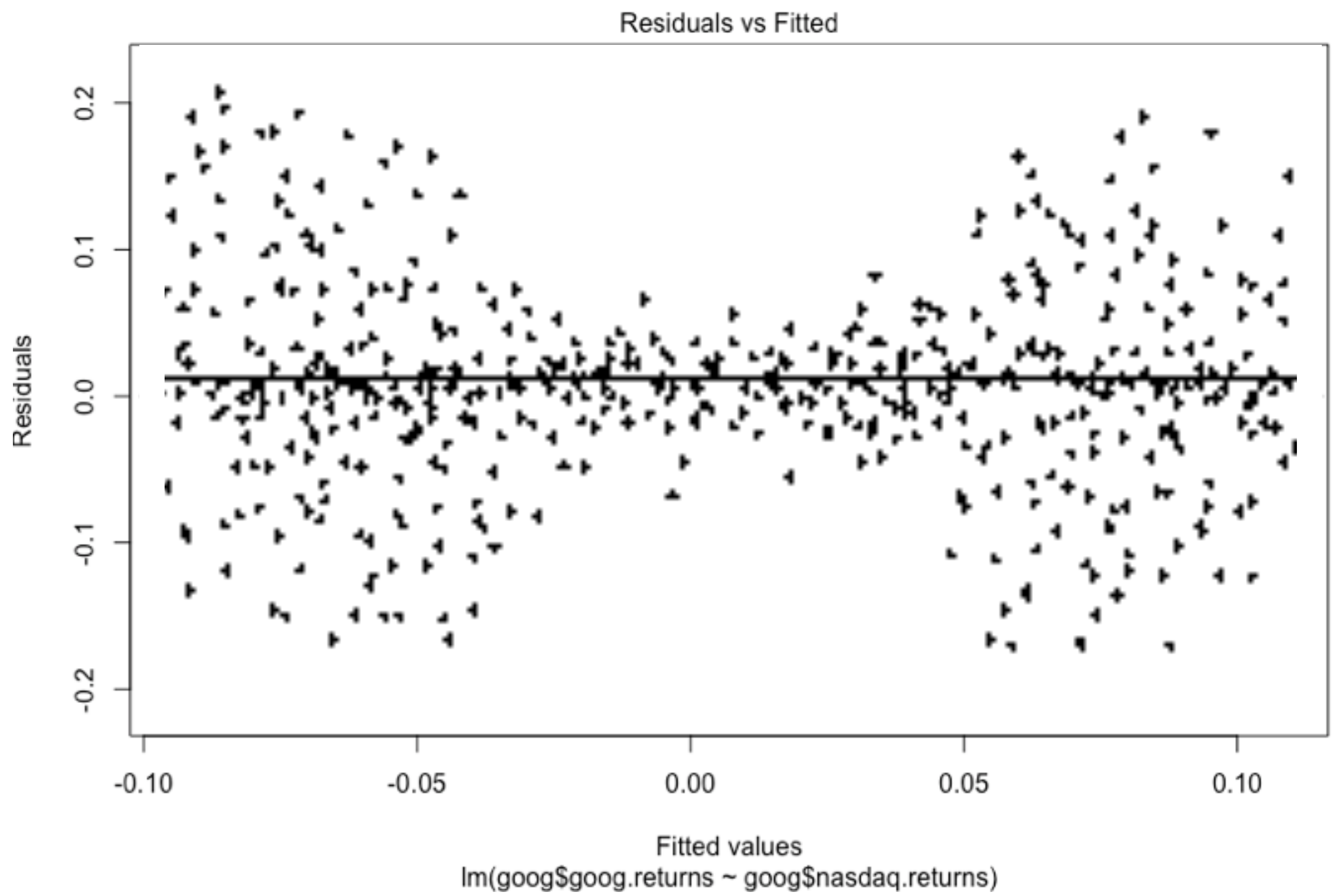
THESE PLOTS REASONABLY SUPPORT THE LINEAR REGRESSION ASSUMPTION

ASSUMPTION 2:

THE VARIANCE OF THE RESIDUALS DOES NOT CHANGE

HERE ARE A COUPLE OF EXAMPLES THAT DO NOT SUPPORT THE ASSUMPTION

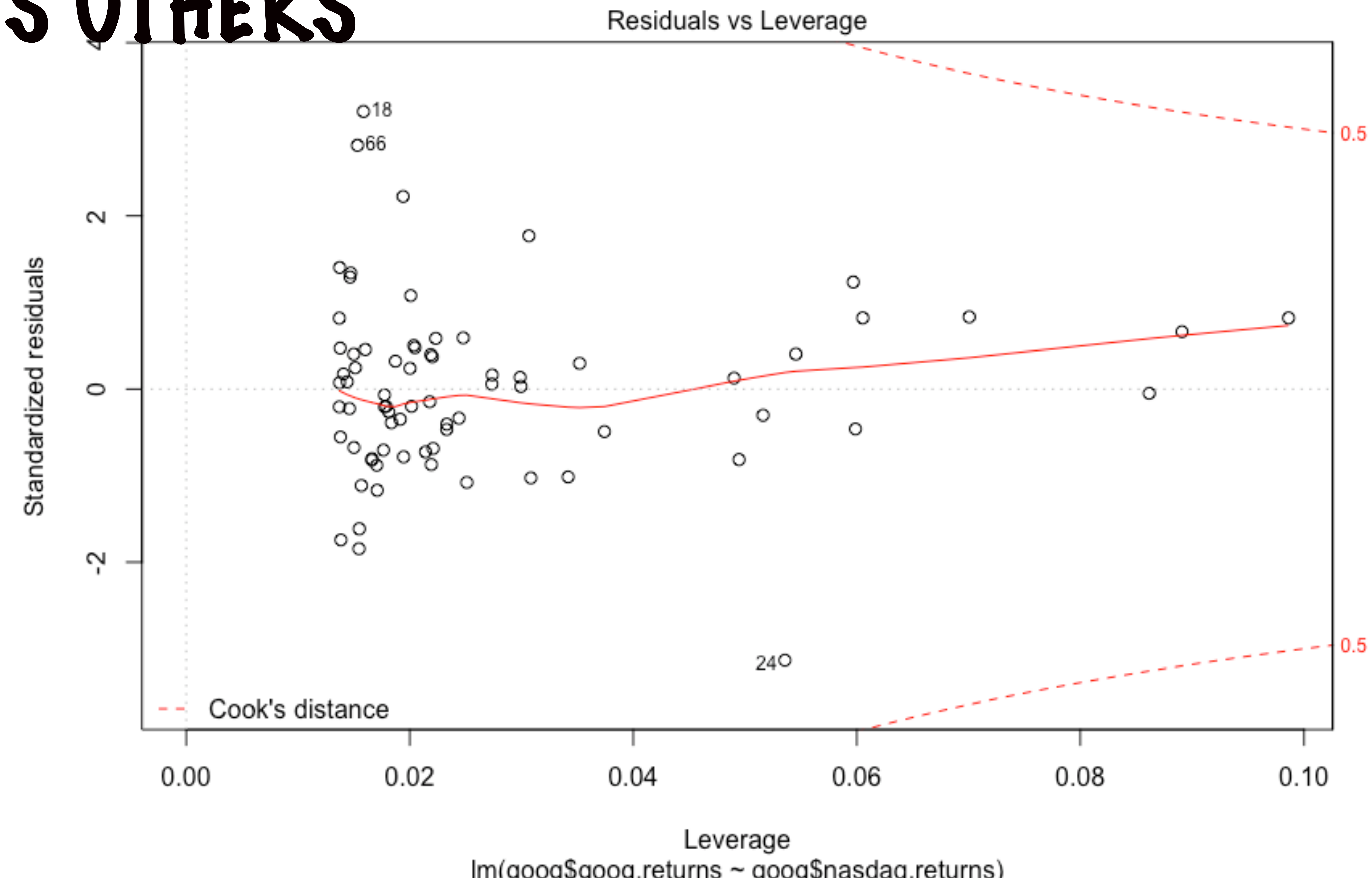




THERE IS ONE MORE PLOT THAT LM() PRINTS

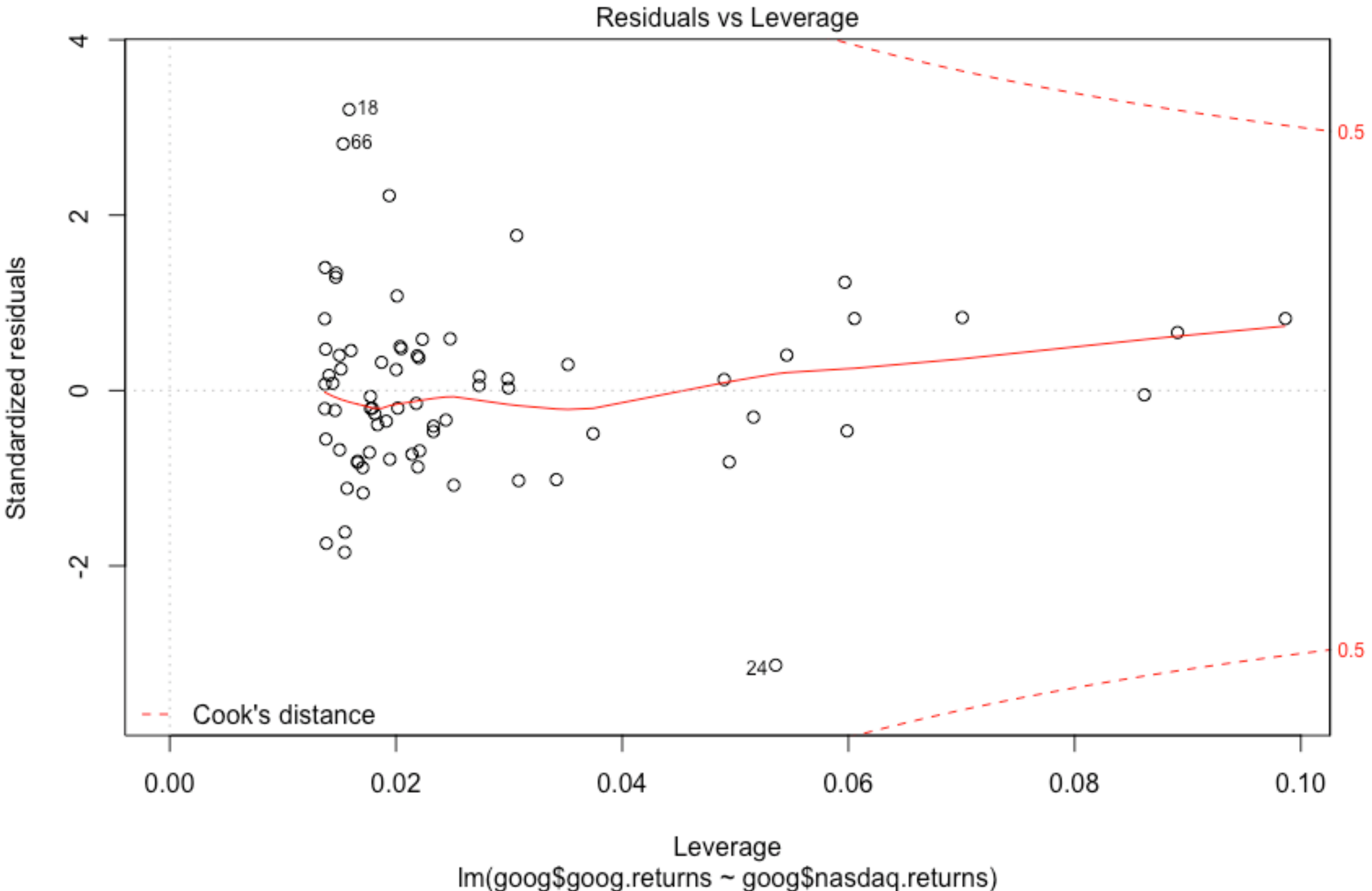
THIS IS NOT TO CHECK AN ASSUMPTION BUT TO SEE IF THERE ARE SOME POINTS (LIKE OUTLIERS) WHICH HAVE MORE INFLUENCE OVER THE REGRESSION RESULT VS OTHERS

COOK'S DISTANCE PLOT



COOK'S DISTANCE COMBINES THESE 2 MEASURES, LEVERAGE AND RESIDUAL VALUE

COOK'S DISTANCE PLOT



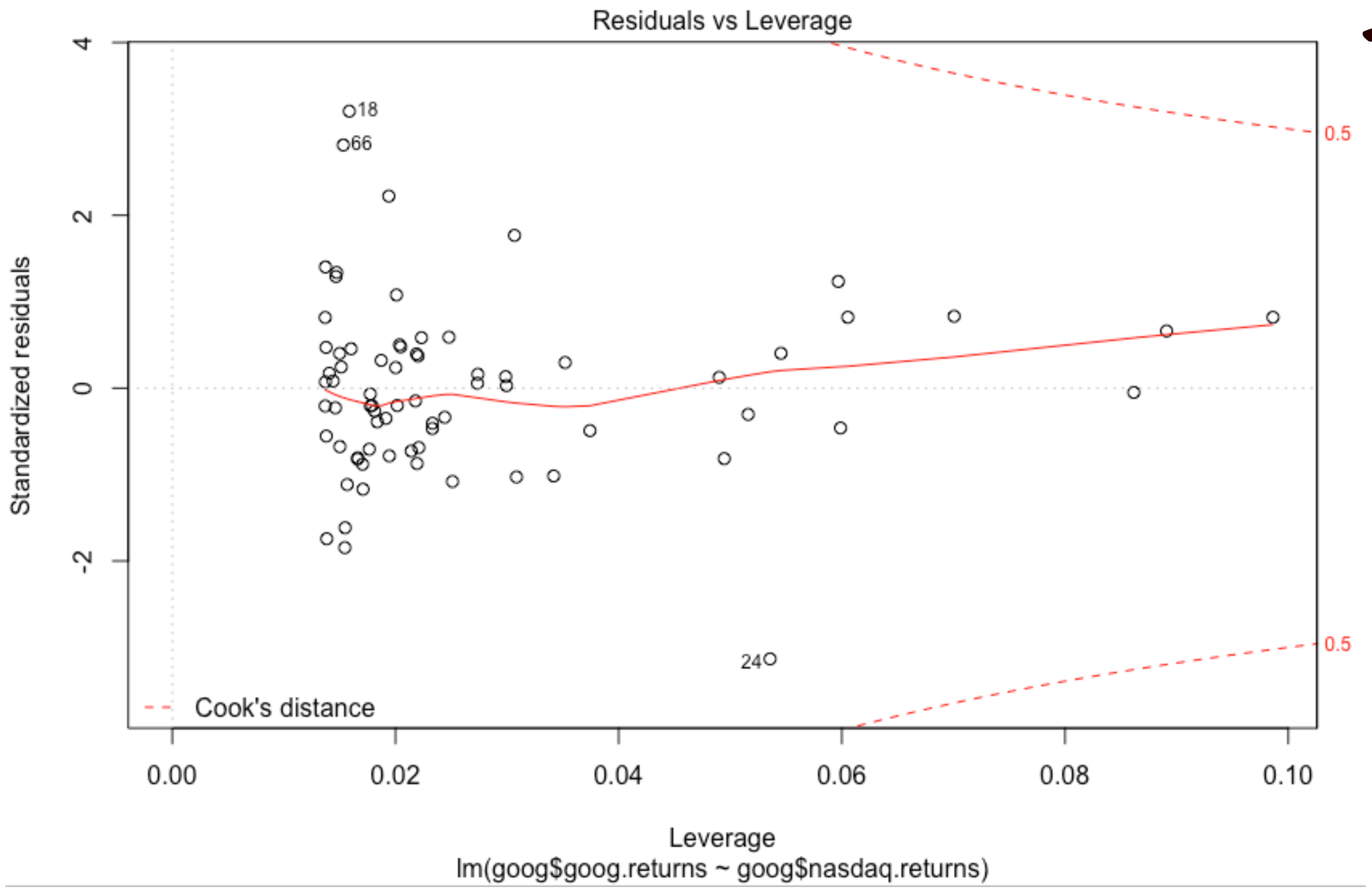
SOME POINTS ARE OUTLIERS, THEY DON'T FOLLOW THE PATTERN OF THE REST OF THE DATA

RESIDUAL VALUES FOR OUTLIERS WILL BE VERY HIGH

IF THE Y-VALUE OF A POINT CHANGES, THE CO-EFFICIENTS WILL CHANGE A LOT

THESE POINTS ARE SAID TO HAVE HIGH LEVERAGE

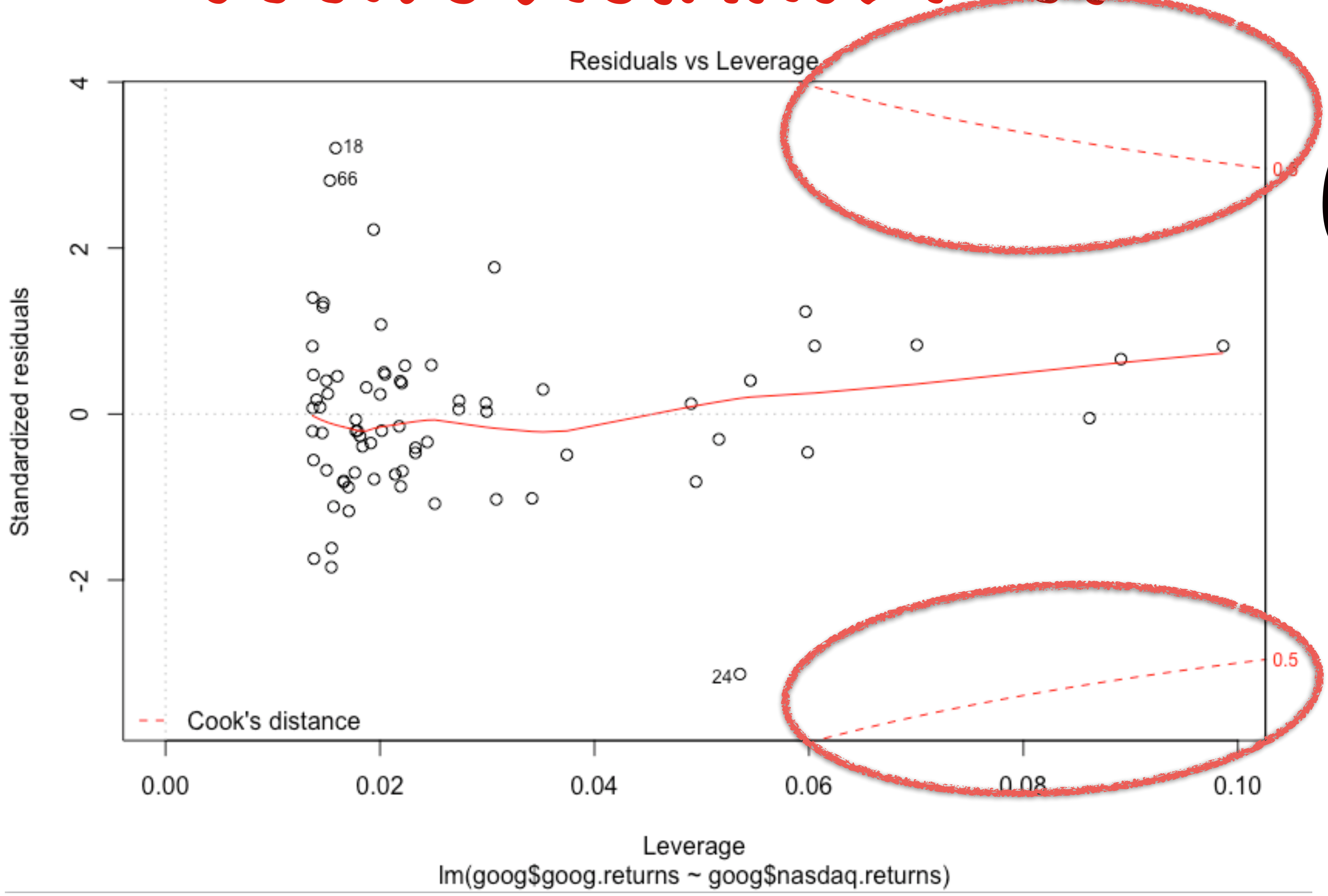
COOK'S DISTANCE PLOT



COOK'S DISTANCE COMBINES THESE 2 MEASURES, LEVERAGE AND RESIDUAL VALUE

THE IDEA BEHIND THIS PLOT IS TO BE AWARE OF POINTS WITH A LARGE COOK'S DISTANCE

COOK'S DISTANCE PLOT



CHECK IF ANY POINTS LIE HERE

THESE POINTS MIGHT NEED FURTHER INVESTIGATION

WAS THERE A DATA ENTRY ERROR? WAS THERE SOME KIND OF EVENT THERE?