Non Normal Dependant Variable with Normally Distributed Residuals

Anas Farah - anasfarah@cmu.edu

24 February, 2020

Amongst the common confusions regarding regression is that either X or Y need to be normally distributed. For the the case of X, it's easier to figure out why that shouldn't be the case (other than it's not a necessary assumption for OLS). You can't have factor variables (sex, social class, etc) if X needed to be normally distributed.

It's slightly less clear from an application stand point why the Y doesn't have to be normally distributed. We know that one of the commonly mentioned regression assumptions is that the error terms of our model need to be normally distributed. The assumptions isn't necessary for the OLS estimator to be BLUE, however it's important when constructing the confidence intervals of our regression coefficients.

Now the question that never crossed my mind was this: **If the error terms of Y are normally distributed is the Y necessarly normally distributed?**

Fortunately, more curious people asked that question. The answer is no.

The example below uses a Y with a bimodal distribution (hence not normal) to showcase how you can have non normally distributed Y with normally distributed residuals.

Simulation

Let's start by simulating 10000 observations from a binomial distribution (in the case below, a bernoulli distribution)

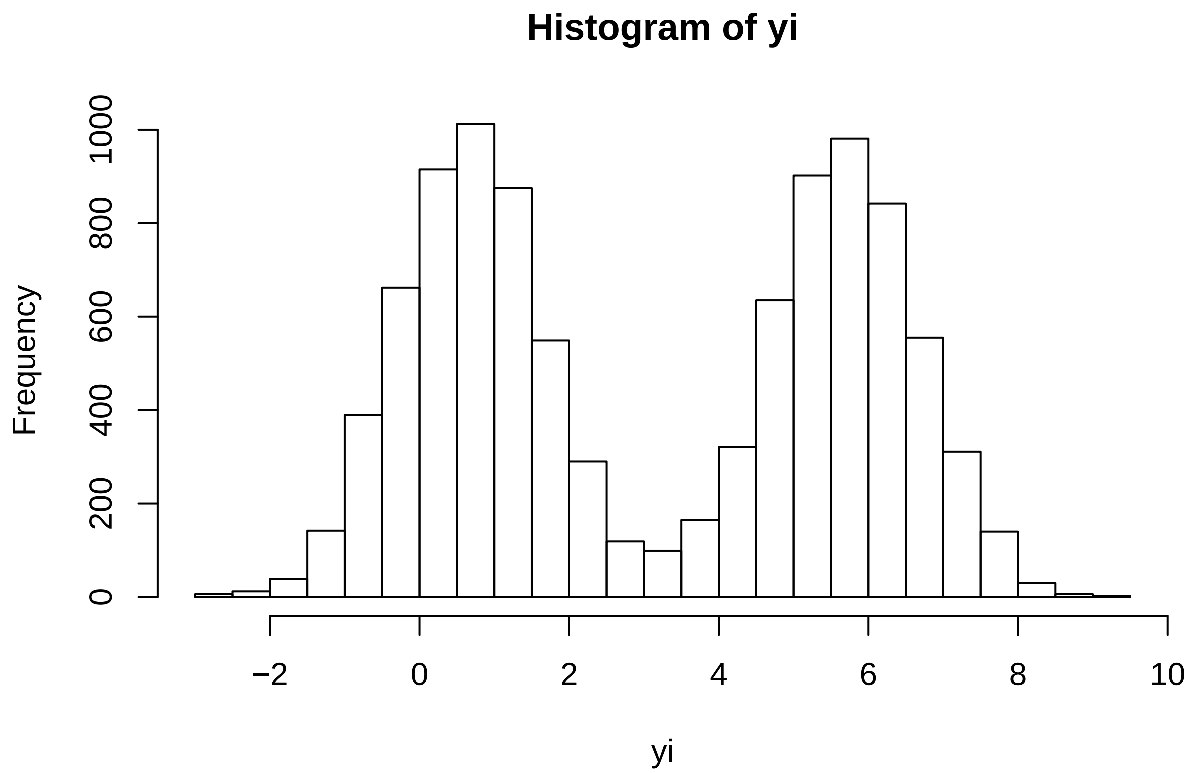
```r
set.seed(1994)
xi <- rbinom(10000, 1, .5)
```

We create a y_i, our dependant variable, from x_i

```r
yi <- 0 + 5 * xi + rnorm(10000, .7)
```

Let's plot a histogram of y_i to visually inspect the distribution of y_i

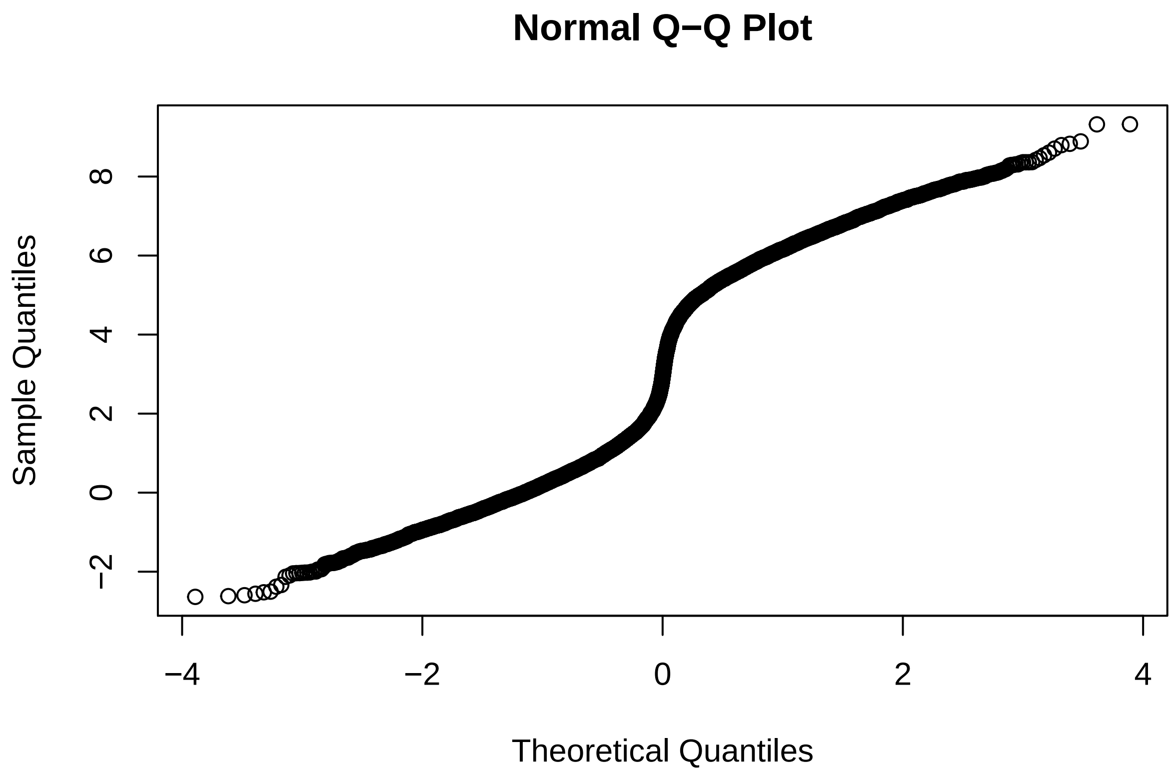
```r
hist(yi, breaks=20)
```



As we can see from the histogram above, y_i has a bimodal distribution (since our x_i is binomial, dichotomous). We have achieved our goal of creating a y_i that's not normally distributed.

We can also look at the qq plot of the y_i to double check the distribution again.

```
qqnorm(yi)
```



From the QQ plot above we can see that the y_i dependant variable is not normally distributed.

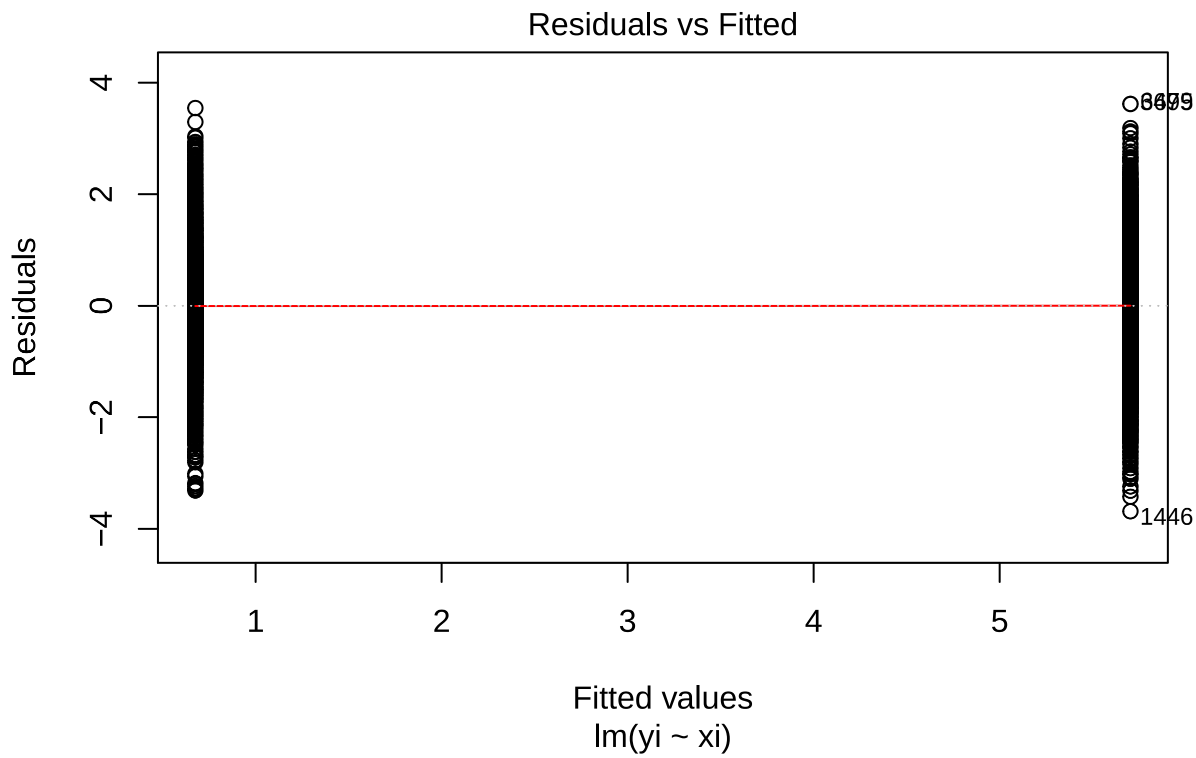
Fitting Our Model

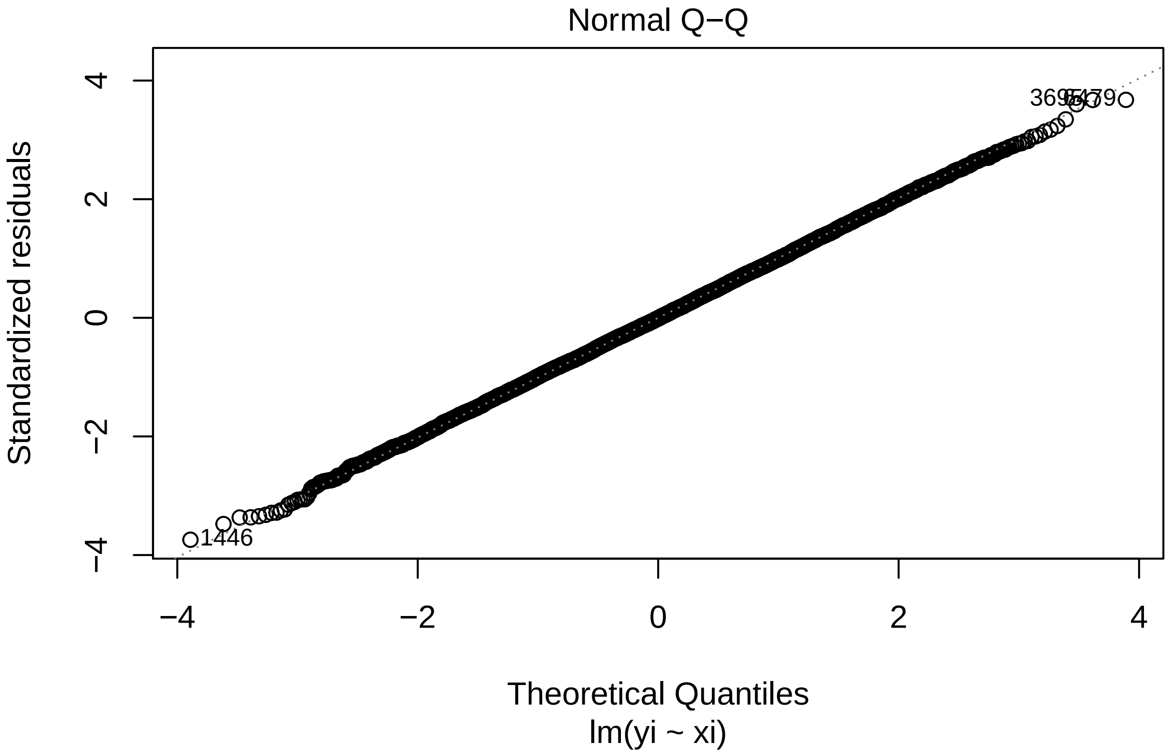
We will now fit a regression of y_i on x_i

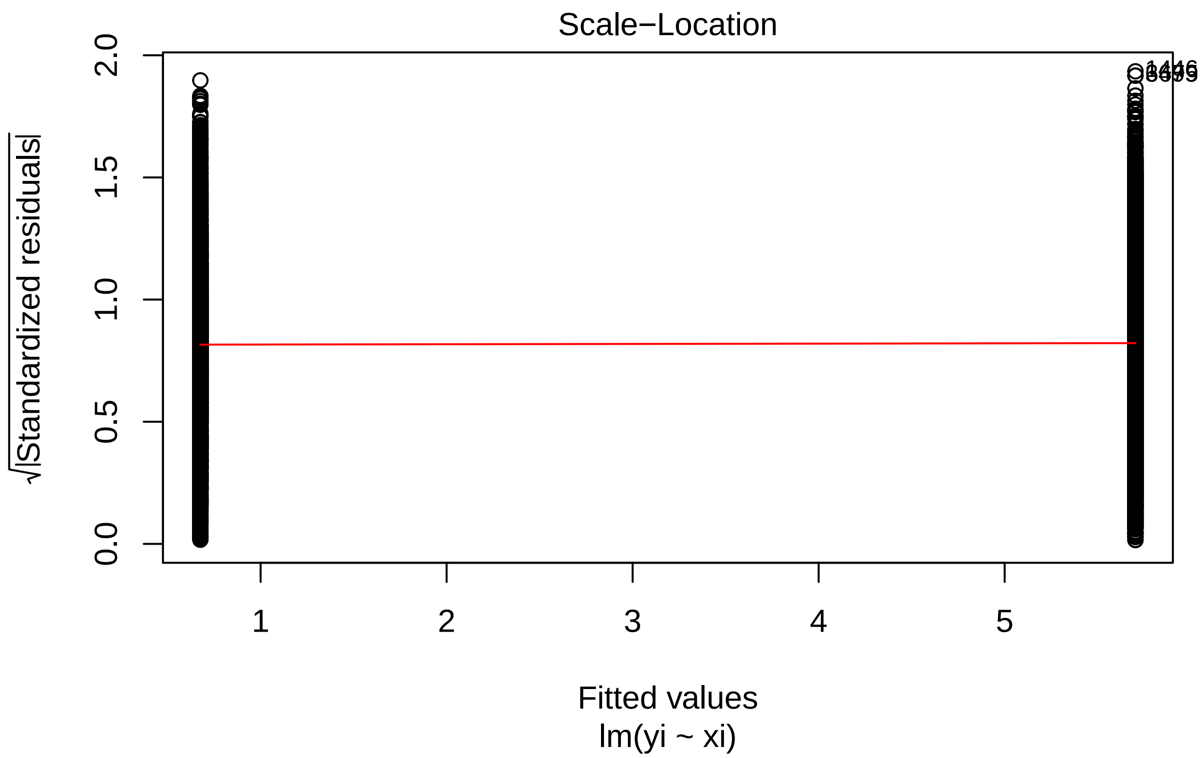
```
model <- lm(yi~xi)
```

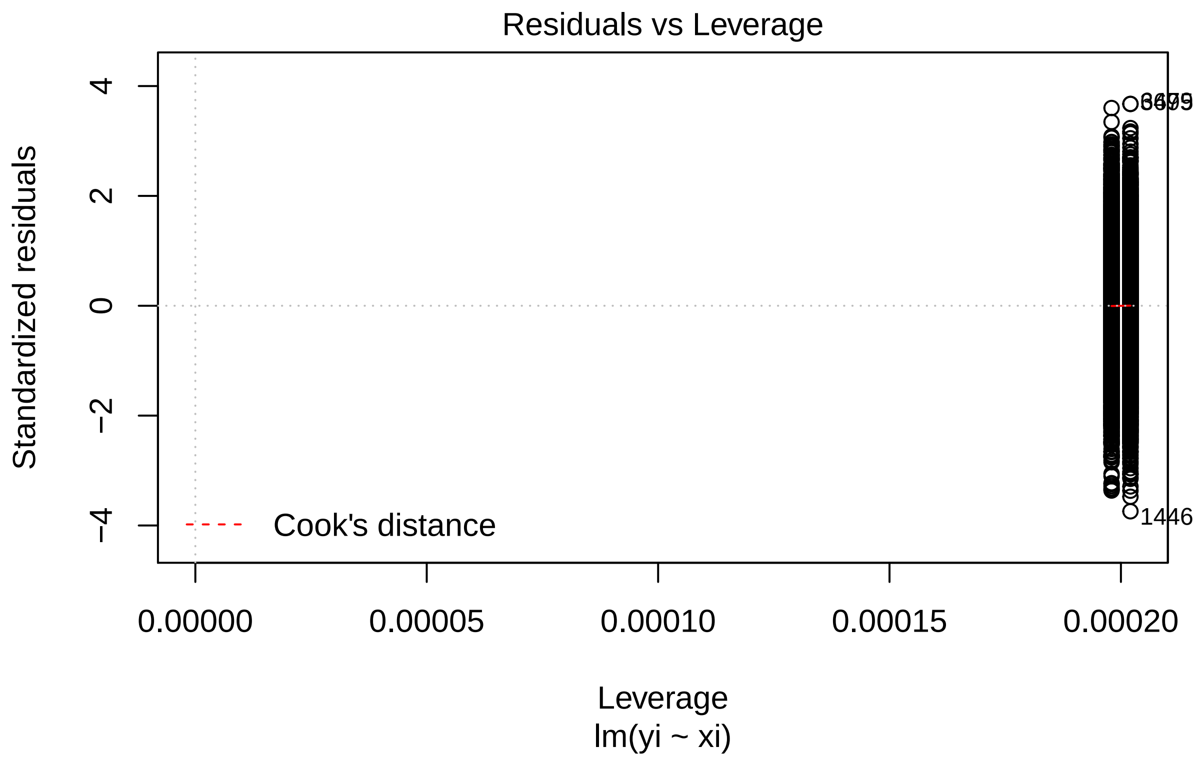
Let's check the diagnostic plots of our regression

```
plot(model)
```


Residuals vs Fitted
Residuals
4
2
0
−2
−4
1
2
3
4
5
1446
Fitted values
lm(yi ~ xi)

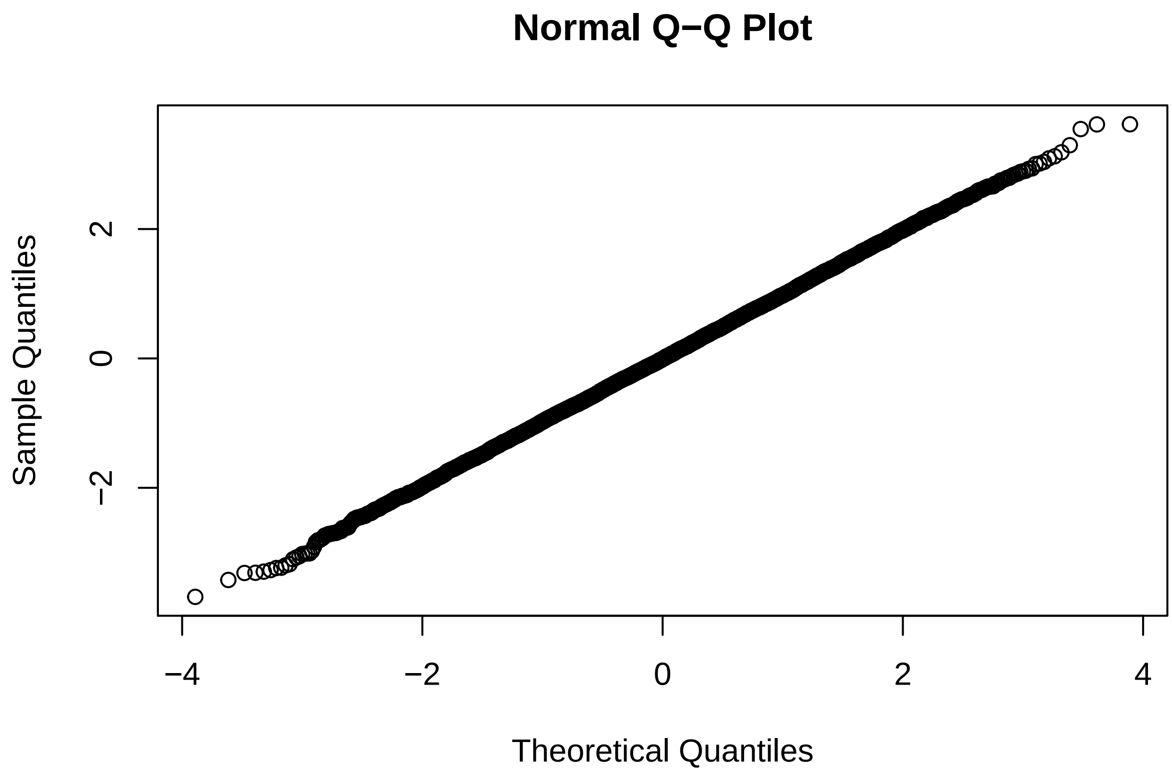

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(yi ~ xi)
1446
4
2
0
−2
−4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
Fitted values
lm(yi ~ xi)



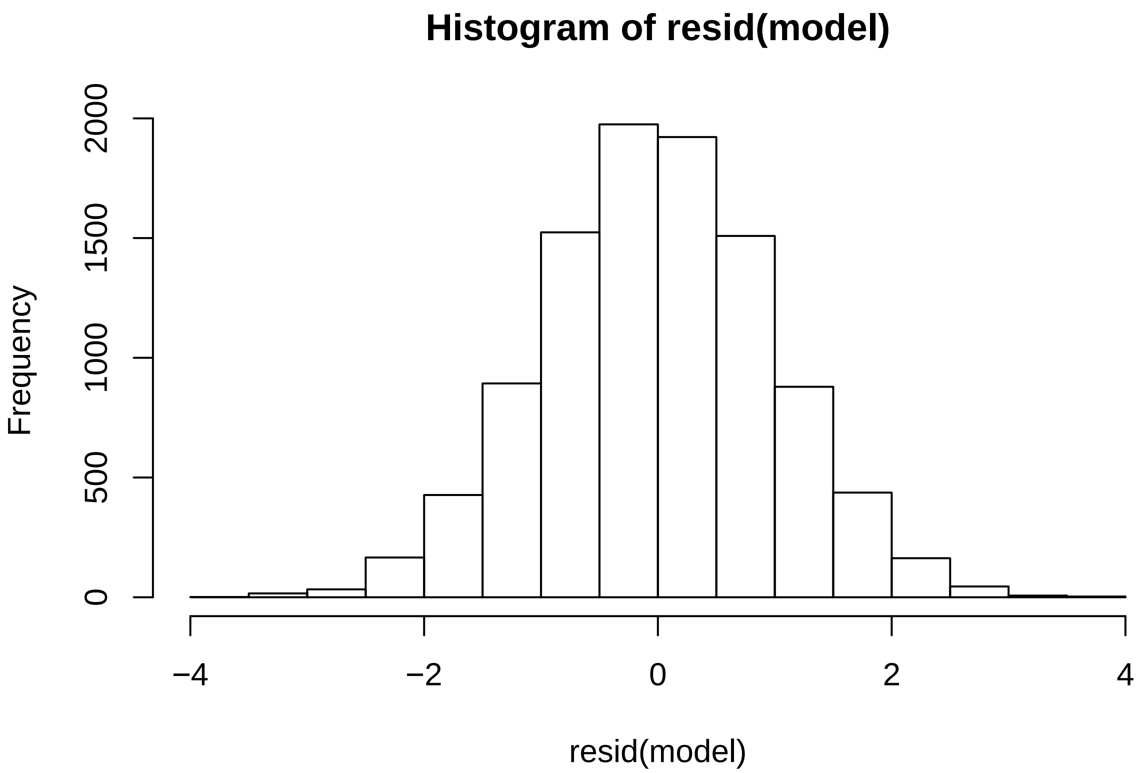
The residuals plot we get is a consequence of using a bimodal x_i. But our focus is not on the residuals plot but on the distribution of the residuals which we can see in the 2nd plot (QQ plot). Let's just focus on that one below

```
qqnorm(resid(model))
```



As we see our residuals are normally distributed even though our y_i dependant variable wasn't. We can check the histogram of our residuals as well to double check

```
hist(resid(model),breaks=20)
```



Conclusion

We have shown in this post that normally distributed errors don't require or originate from normally distributed dependant variables. We used the case of a dependant variable with a bimodal distribution and found that it's error terms are normally distributed.

Reference:

http://www.programmingr.com/examples/neat-tricks/sample-r-function/r-rbinom/ Simulating Binomial and Bernoulli distributions in R

https://stats.stackexchange.com/questions/11351/left-skewed-vs-symmetric-distribution-observed/11352#11352 The code above is inspired from this stackexchange post

https://stats.stackexchange.com/questions/12262/what-if-residuals-are-normally-distributed-but-y-is-not Another example using a multimodal distribution of Y