REGRESSION DIAGNOSTICS

TODAY

- Nothing wrong
- Non-normal errors
- Non-constant variance
- Incorrect functional form
- Summary

R

- Simples diagnostics graphs:
  plot(model_name)

- Steps through 4 plots

- Alternative: create residuals, etc, and use ggplot (that is what I have done here; see R code on Canvas)

ASSUMPTIONS

How to examine them

WHAT ASSUMPTIONS?

1. Error term has a zero population mean
2. Error terms not correlated with each other
3. Error term has a constant variance
4. Error term is normally distributed

STANDARDIZED RESIDUALS

$$\frac{(y_i - \hat{y}_i) - \frac{\sum_i^n (y_i - \hat{y}_i)}{n}}{s_\varepsilon}$$

Why?

Once standardized it is easy
to check for normality

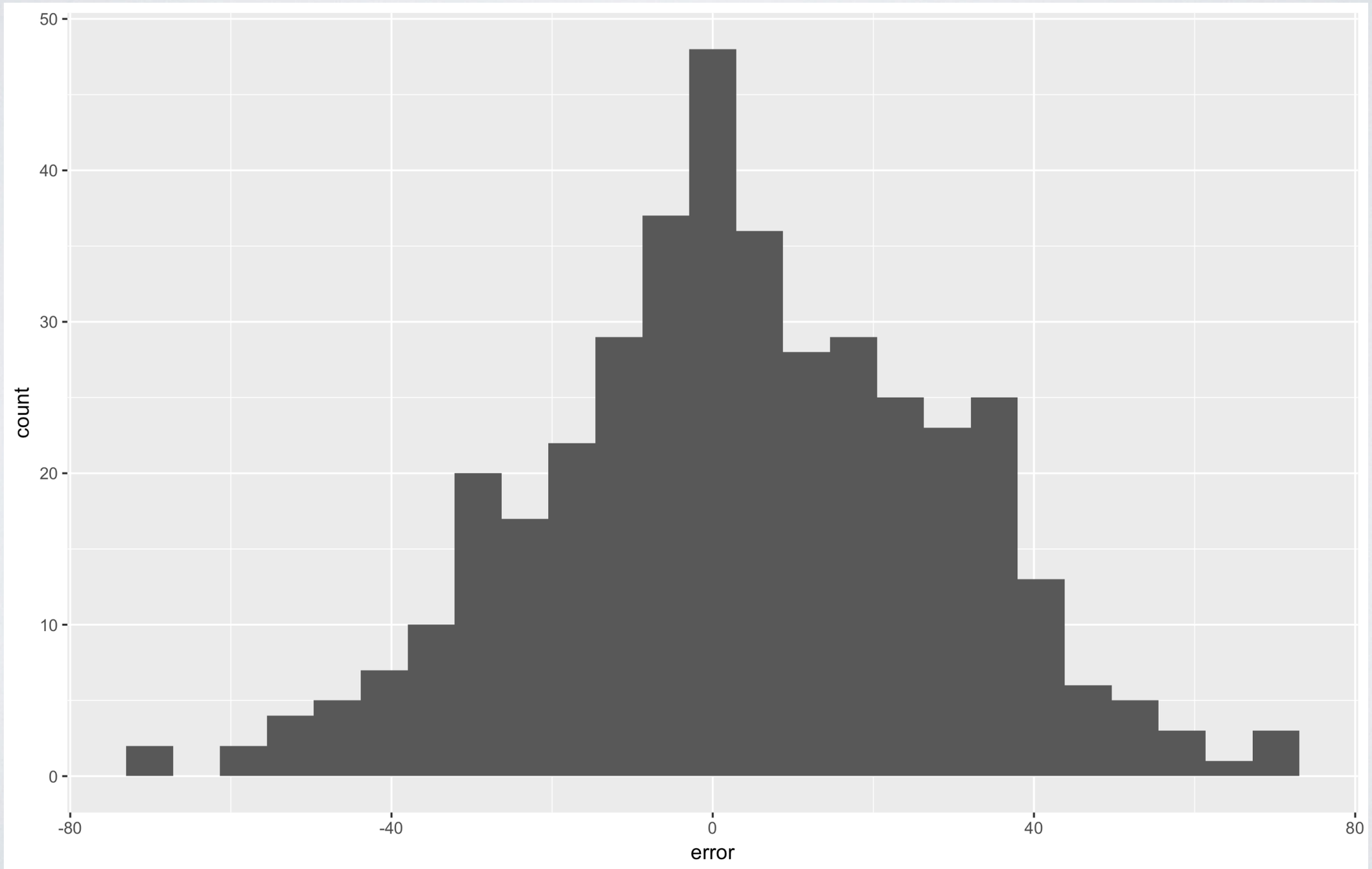
EXAMPLE WHEN ASSUMPTIONS HOLD

CODE

```
set.seed(2)
x <- runif(400) * 100
error <- rnorm(400, 0, 25)
df <- data.frame(x, error) %>%
  mutate(y = 1.5 * x + error + 40)
```

MADE-UP DATA

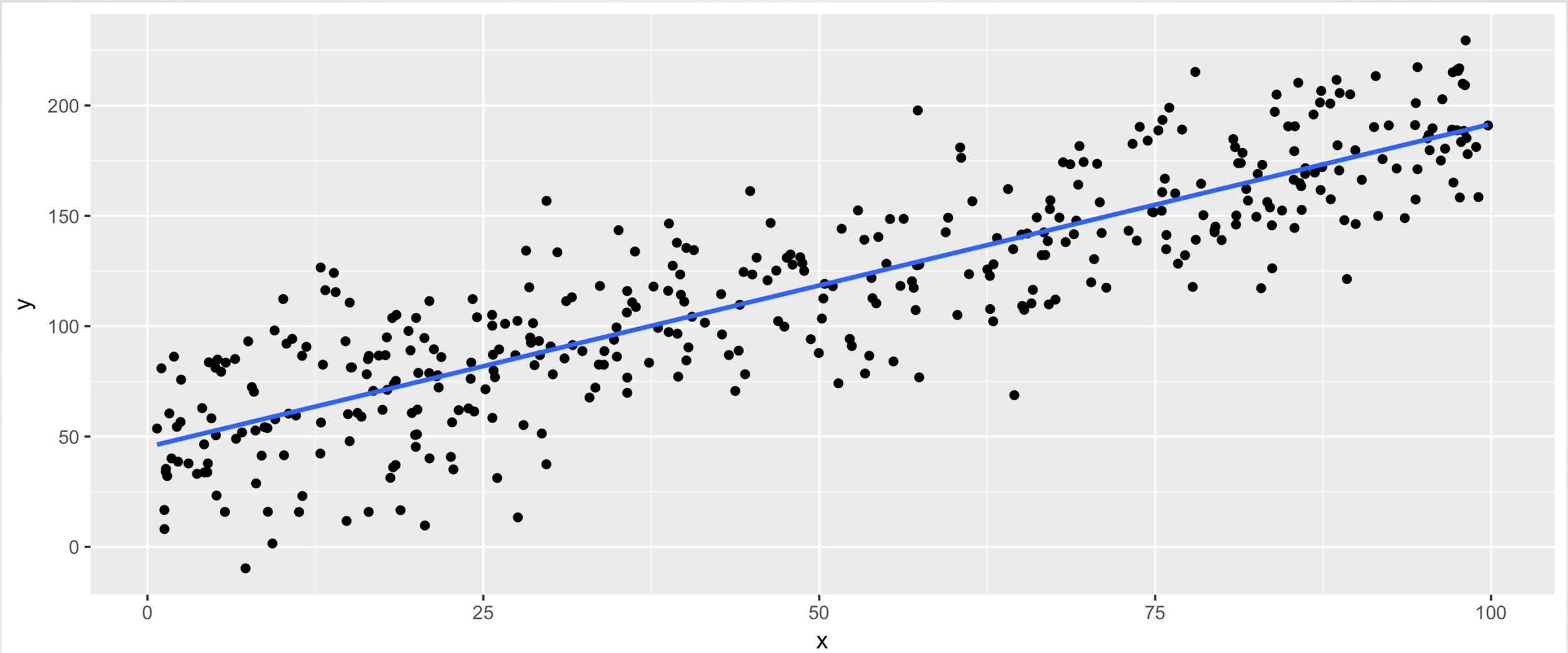
$$Y = 40 + 1.5X + \text{normal error}$$



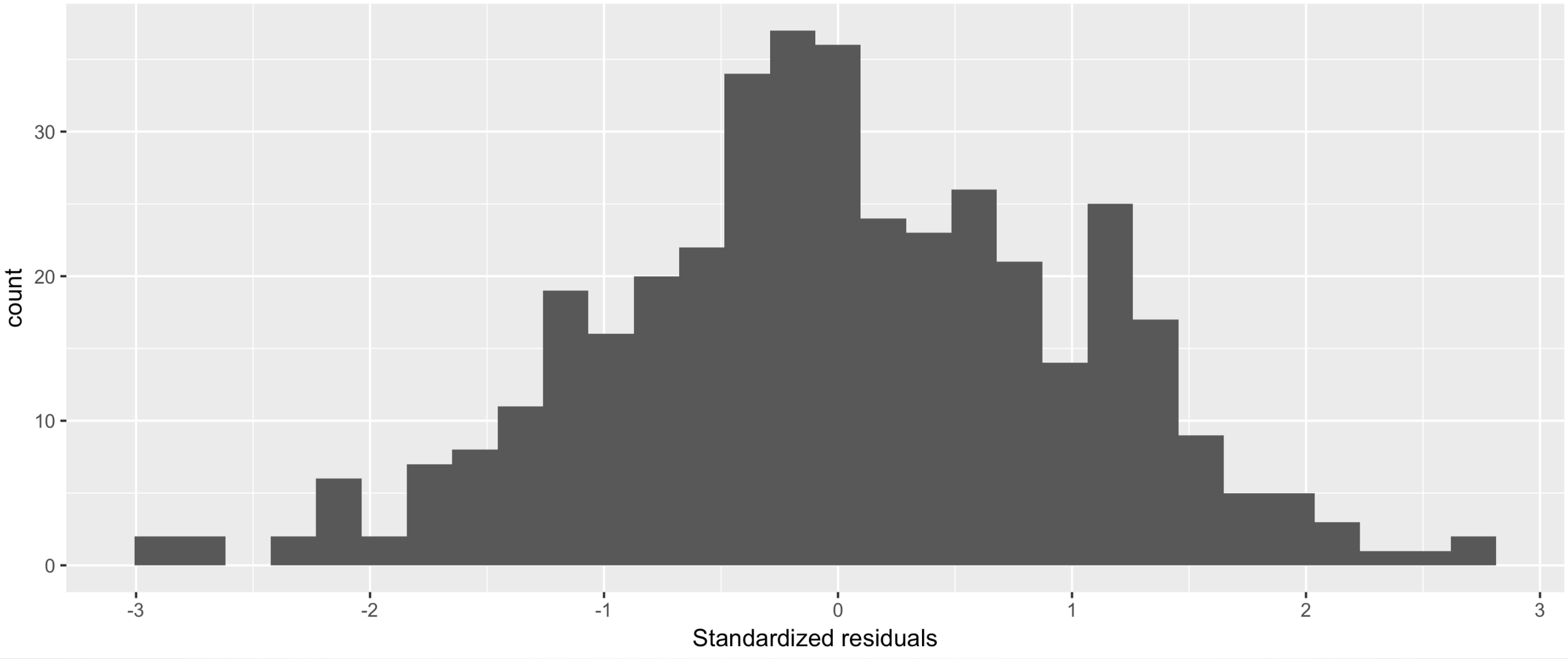
RESULTS

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-72.281 -16.410  -0.788  17.364  68.589

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.28968    2.36508   19.15   <2e-16 ***
x            1.46326    0.04151   35.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.08 on 398 degrees of freedom
Multiple R-squared:  0.7574,    Adjusted R-squared:  0.7568
F-statistic:  1242 on 1 and 398 DF,  p-value: < 2.2e-16
```

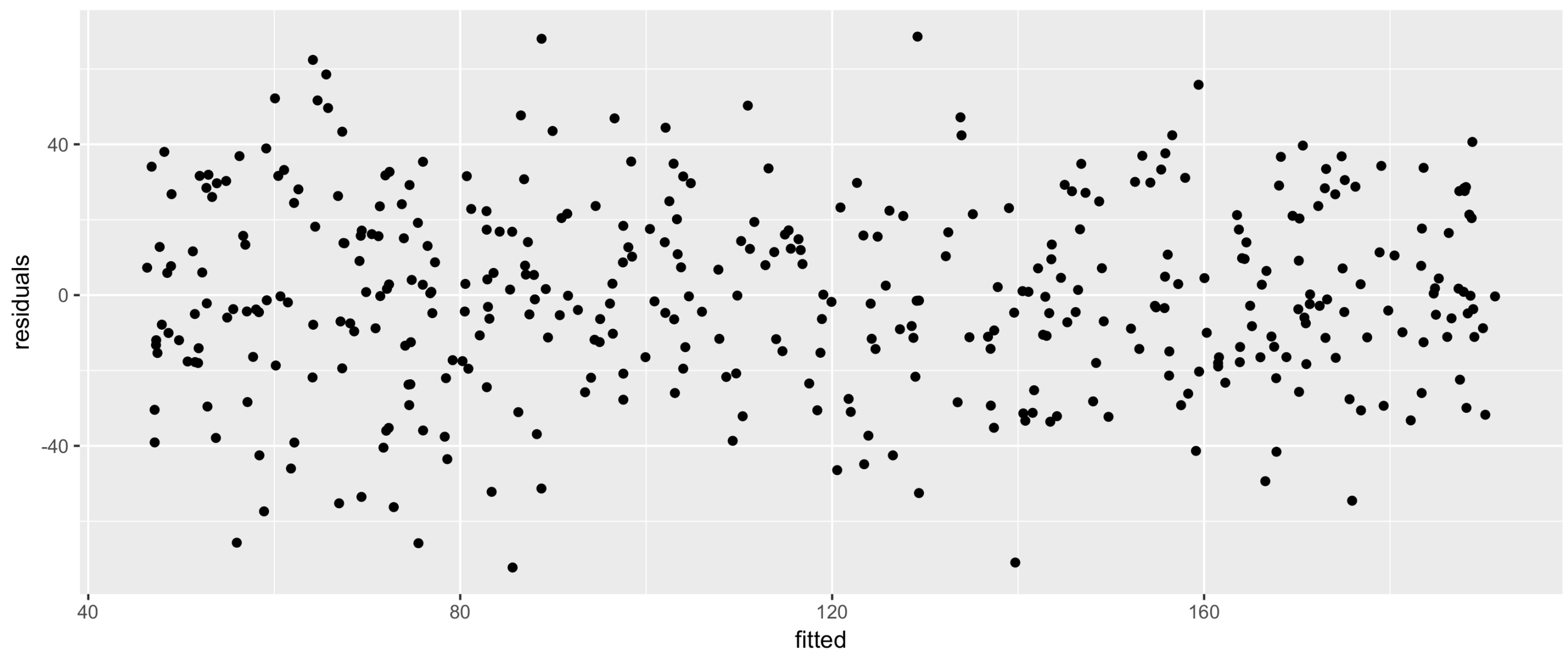
REGRESSION GRAPH



ERRORS NORMAL



CONSTANT VARIANCE?

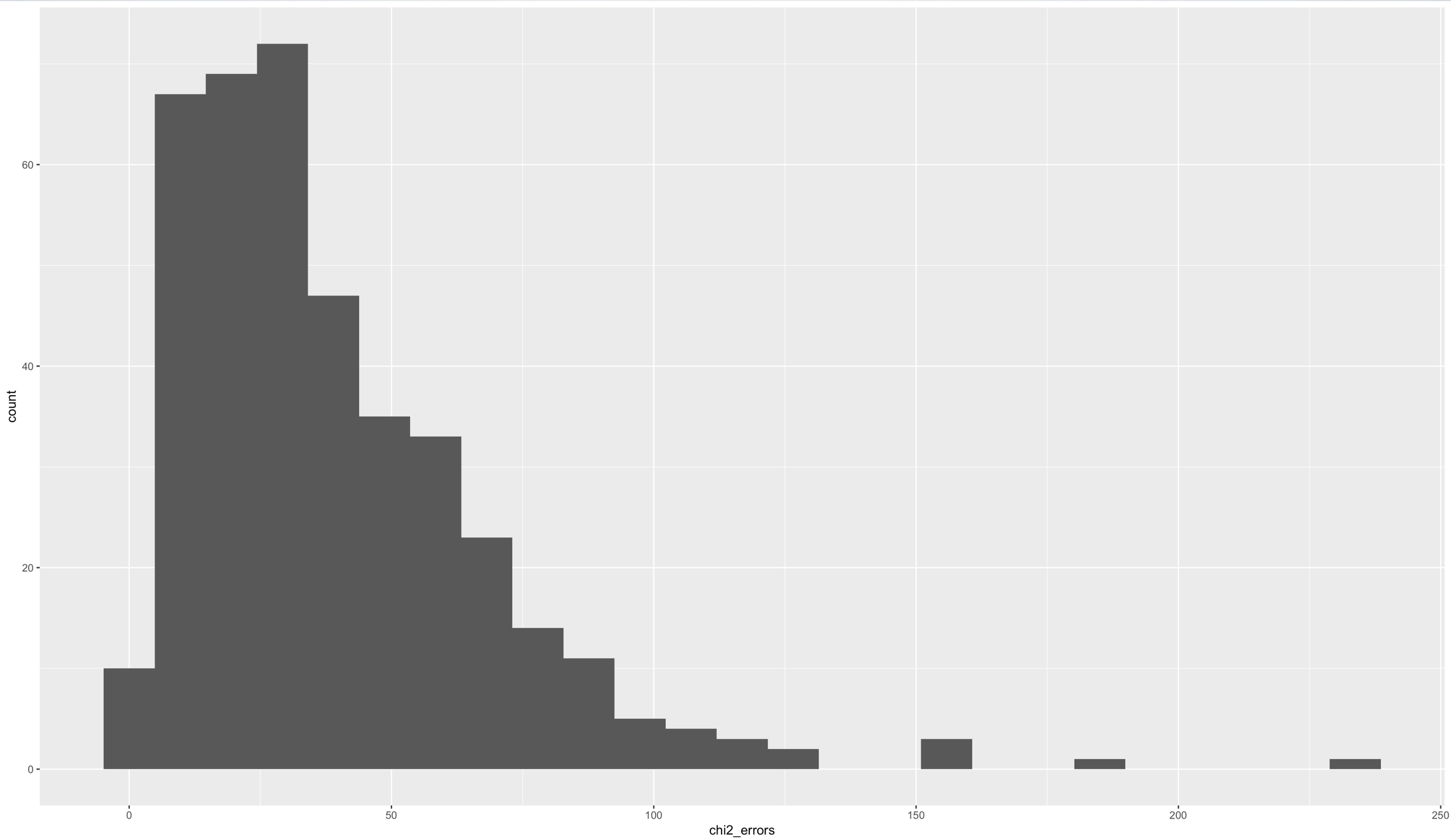


EXAMPLE WHEN ERRORS NOT NORMAL

CODE

```
set.seed(4)
df$chi2_errors <- rchisq(400, 4) * 10
df <- df %>%
  mutate(y_non_normal = 1.5 * x + chi2_errors)
```

CHI^2 DATA



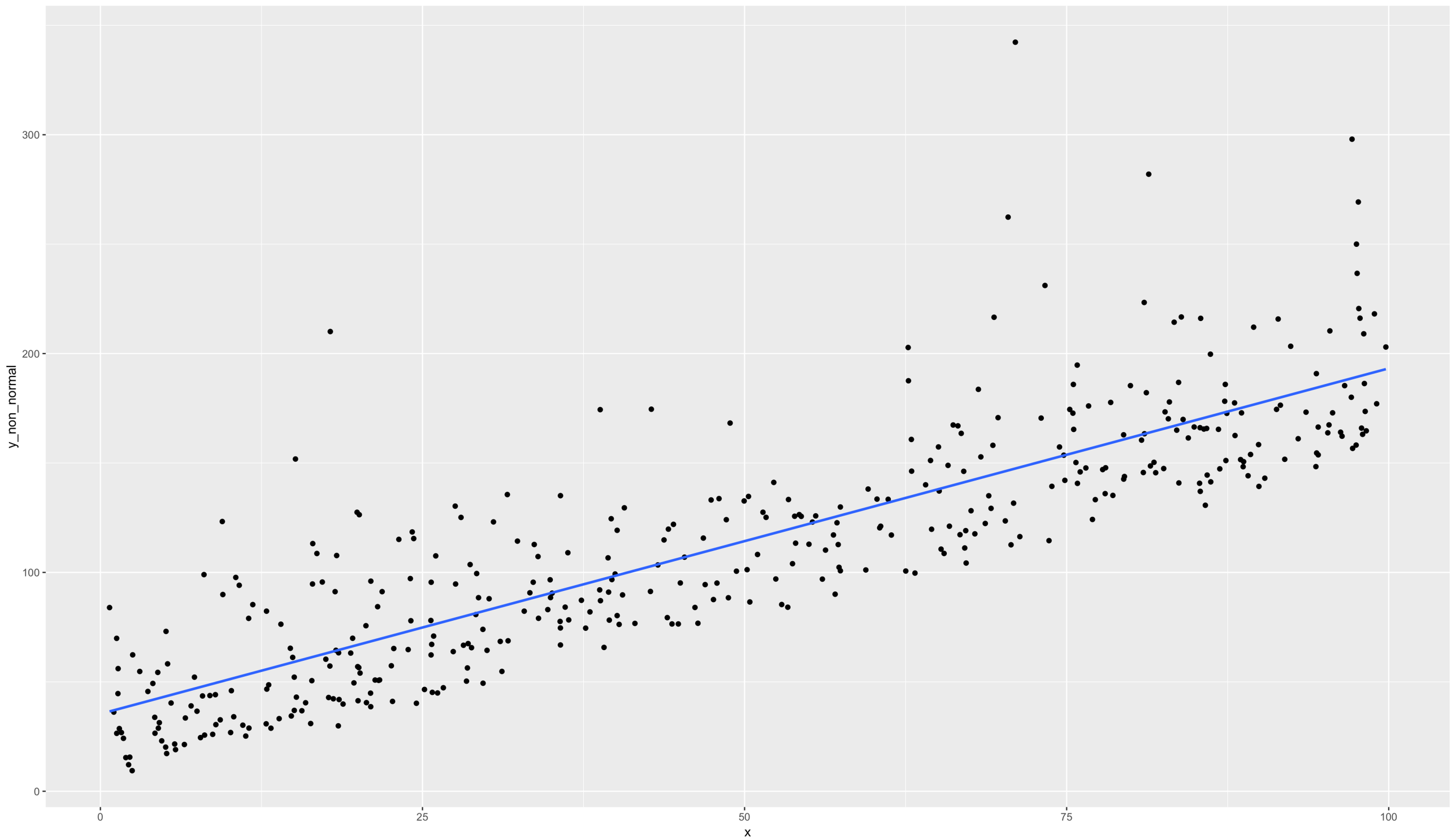
OUR "MODEL"

$Y = 1.5\,X + Chi^2$ error

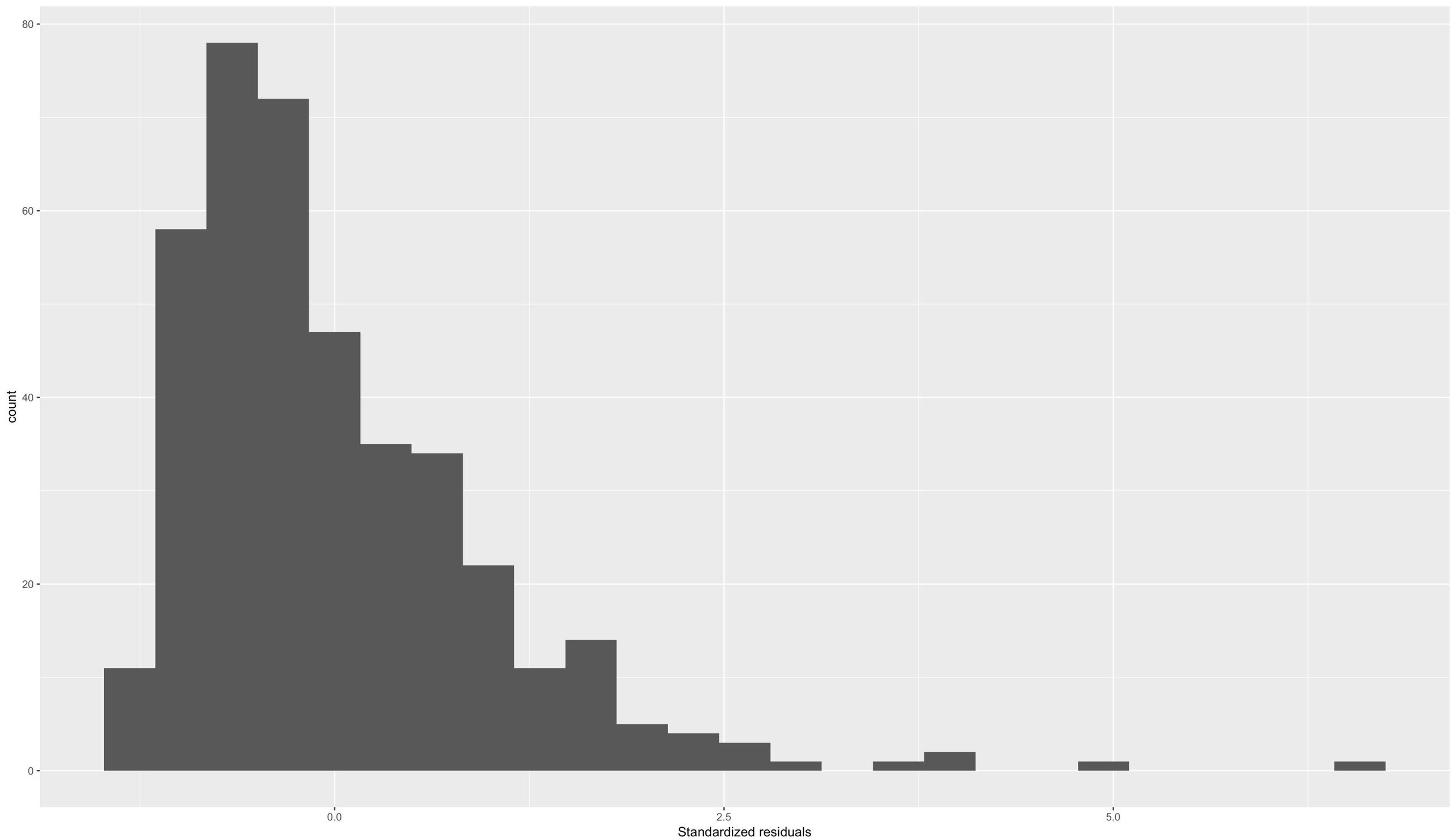
Generate 400 observations

```
Call:
lm(formula = y_non_normal ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-40.074 -22.208  -7.579  14.575 194.803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.36017    2.80915   12.59   <2e-16 ***
x            1.57872    0.04931   32.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.79 on 398 degrees of freedom
Multiple R-squared:  0.7203,    Adjusted R-squared:  0.7196
F-statistic:  1025 on 1 and 398 DF,  p-value: < 2.2e-16
```

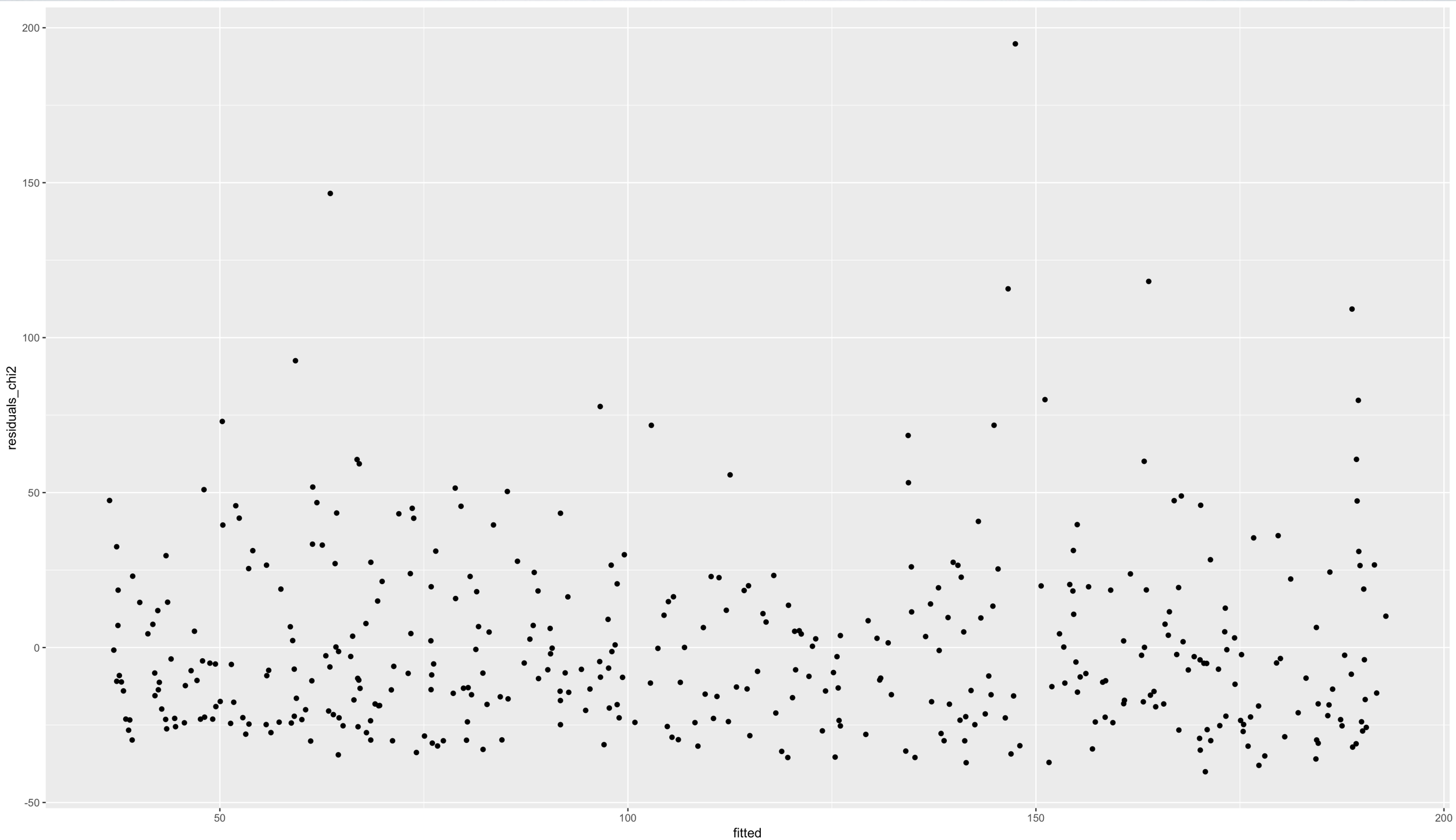
REGRESSION GRAPH



ERRORS NORMAL?



HOMOSKEDASTICITY?



EXAMPLE WITH NON-CONSTANT VARIANCE OF ERRORS

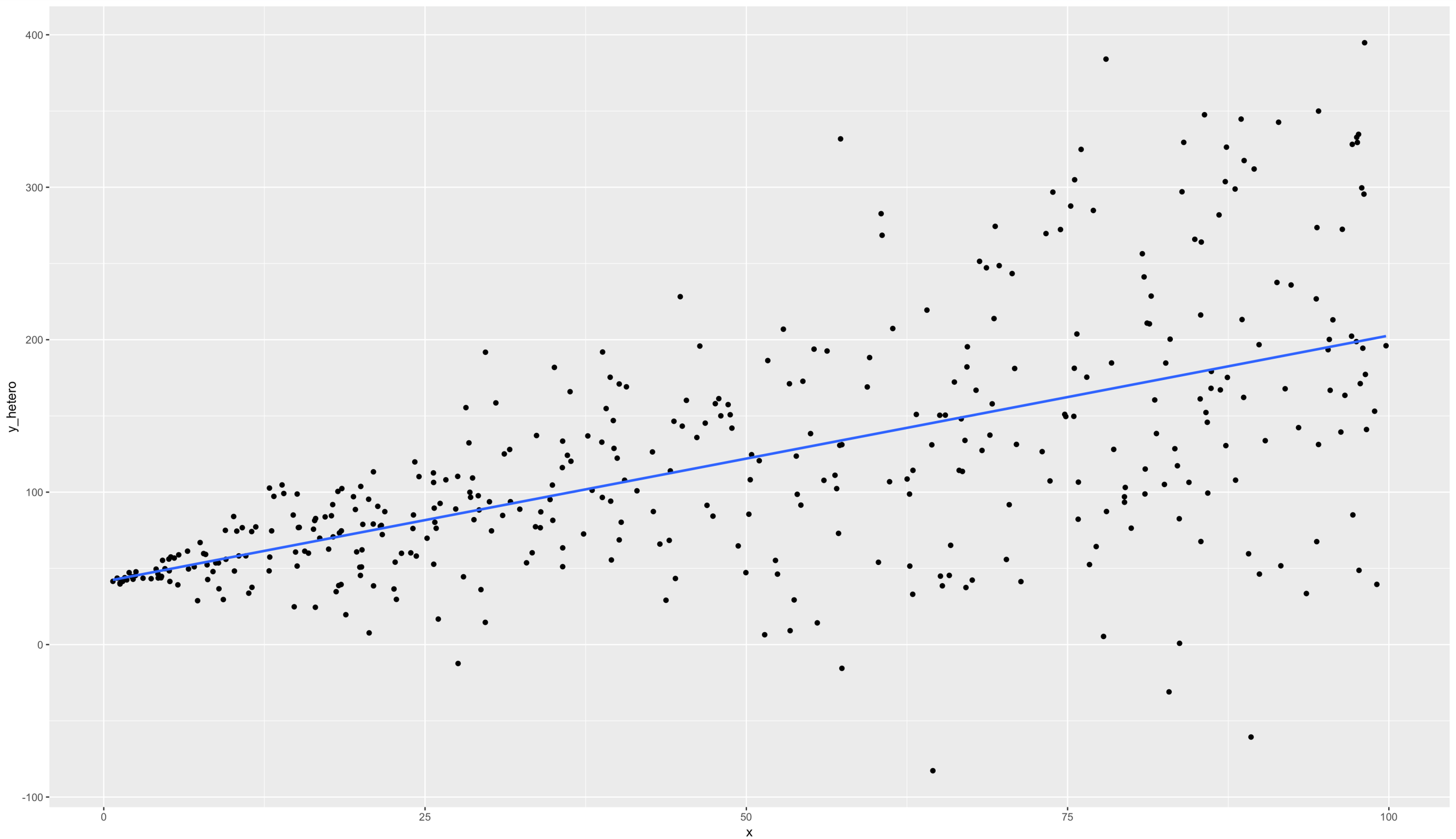
HETEROSKEDASTICITY

```
df <- df %>%
  mutate(
    hetero_error = error * x/20,
    y_hetero = 40 + 1.5 * x + hetero_error
  )
```

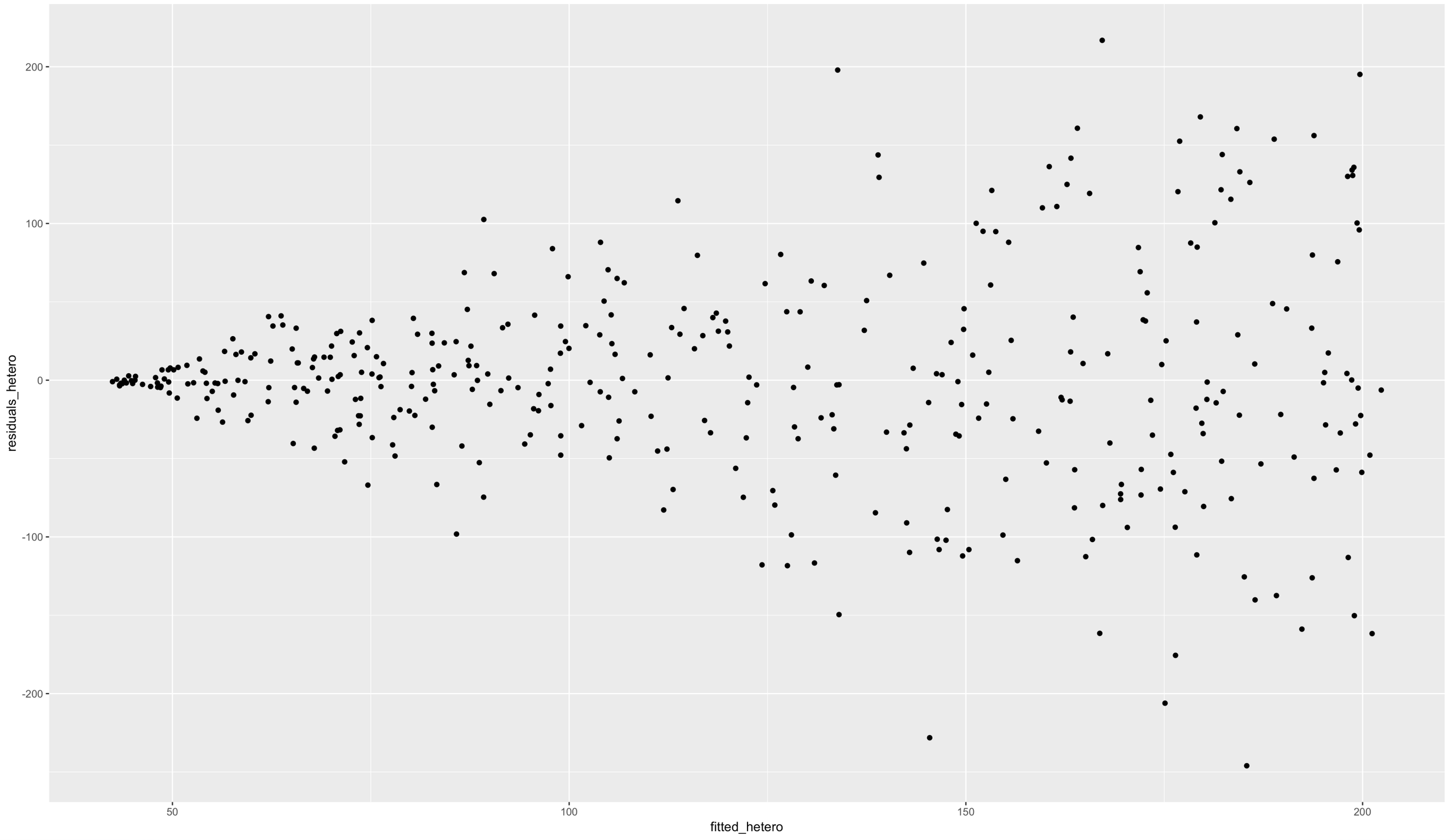
RESULTS

```
Call:
lm(formula = y_hetero ~ x, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-246.038  -31.203   -1.178   29.316  216.868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.3004     6.1927   6.669 8.62e-11 ***
x             1.6141     0.1087  14.850  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.68 on 398 degrees of freedom
Multiple R-squared:  0.3565,    Adjusted R-squared:  0.3549
F-statistic: 220.5 on 1 and 398 DF,  p-value: < 2.2e-16
```

REGRESSION GRAPH



CONSTANT VARIANCE?



EXAMPLE WITH INCORRECT FUNCTIONAL FORM FOR MODEL

CODE

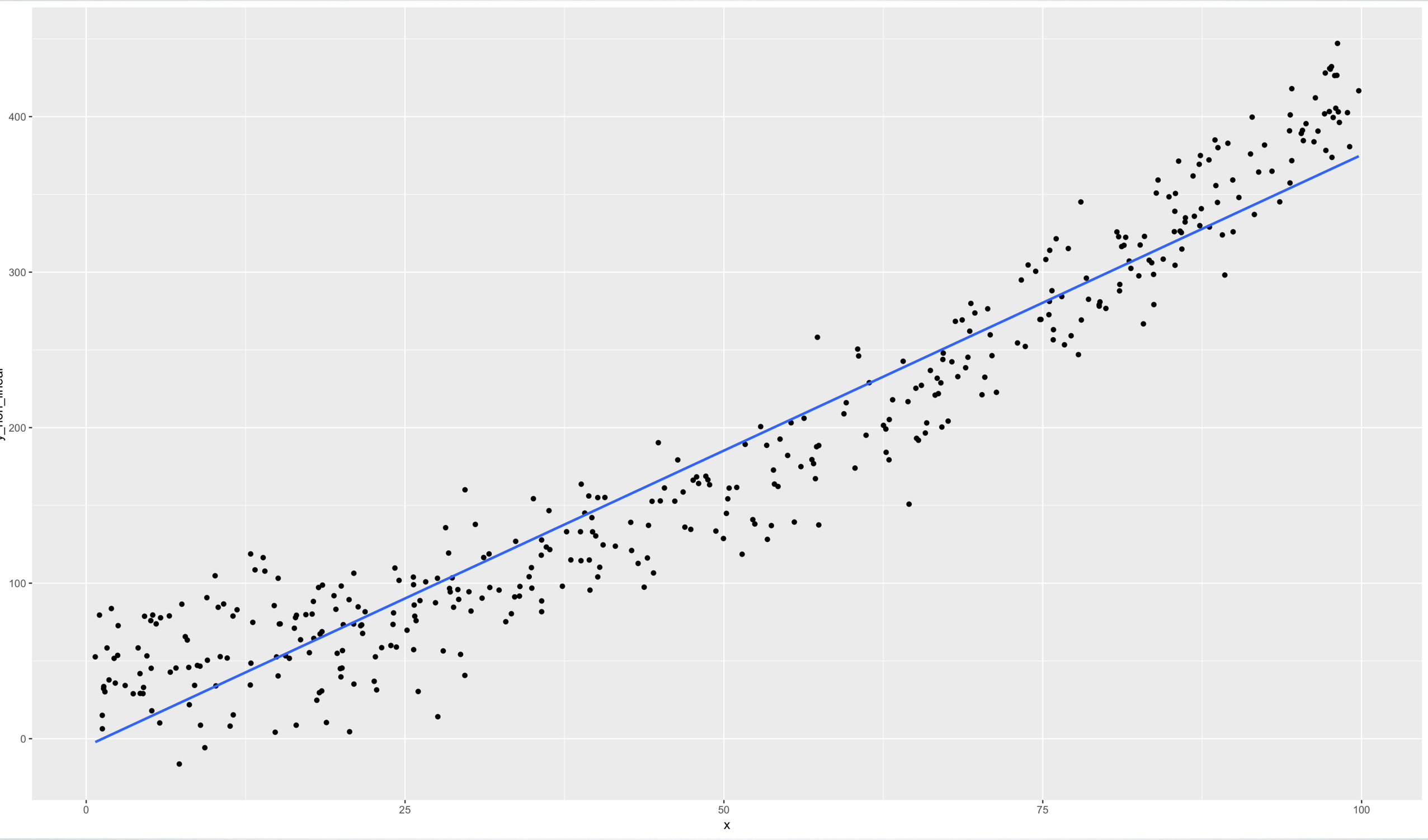
```
df <- df %>%
  mutate(
    y_non_linear = 40 + 1.5 * x^1.7 / 10 + error
  )
```

NON-LINEAR MODEL

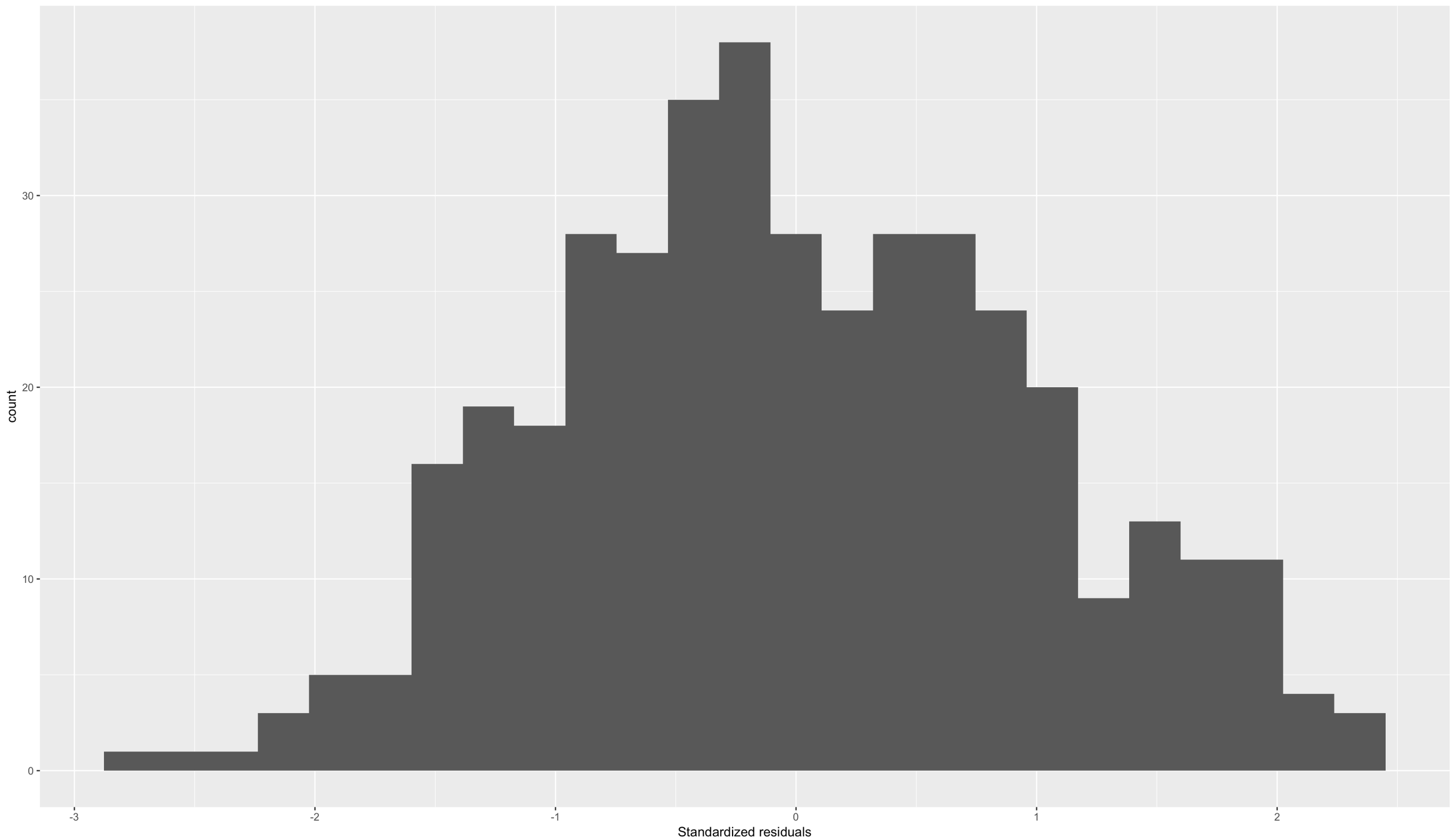
$$Y = 40 + 1.5X^{1.7}/10 + \text{normal error}$$

```
Call:
lm(formula = y_non_linear ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-89.648 -23.908  -2.546  22.705  81.058

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.86670    3.15598  -1.542    0.124
x            3.80270    0.05539  68.648   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.47 on 398 degrees of freedom
Multiple R-squared:  0.9221,    Adjusted R-squared:  0.9219
F-statistic:  4713 on 1 and 398 DF,  p-value: < 2.2e-16
```

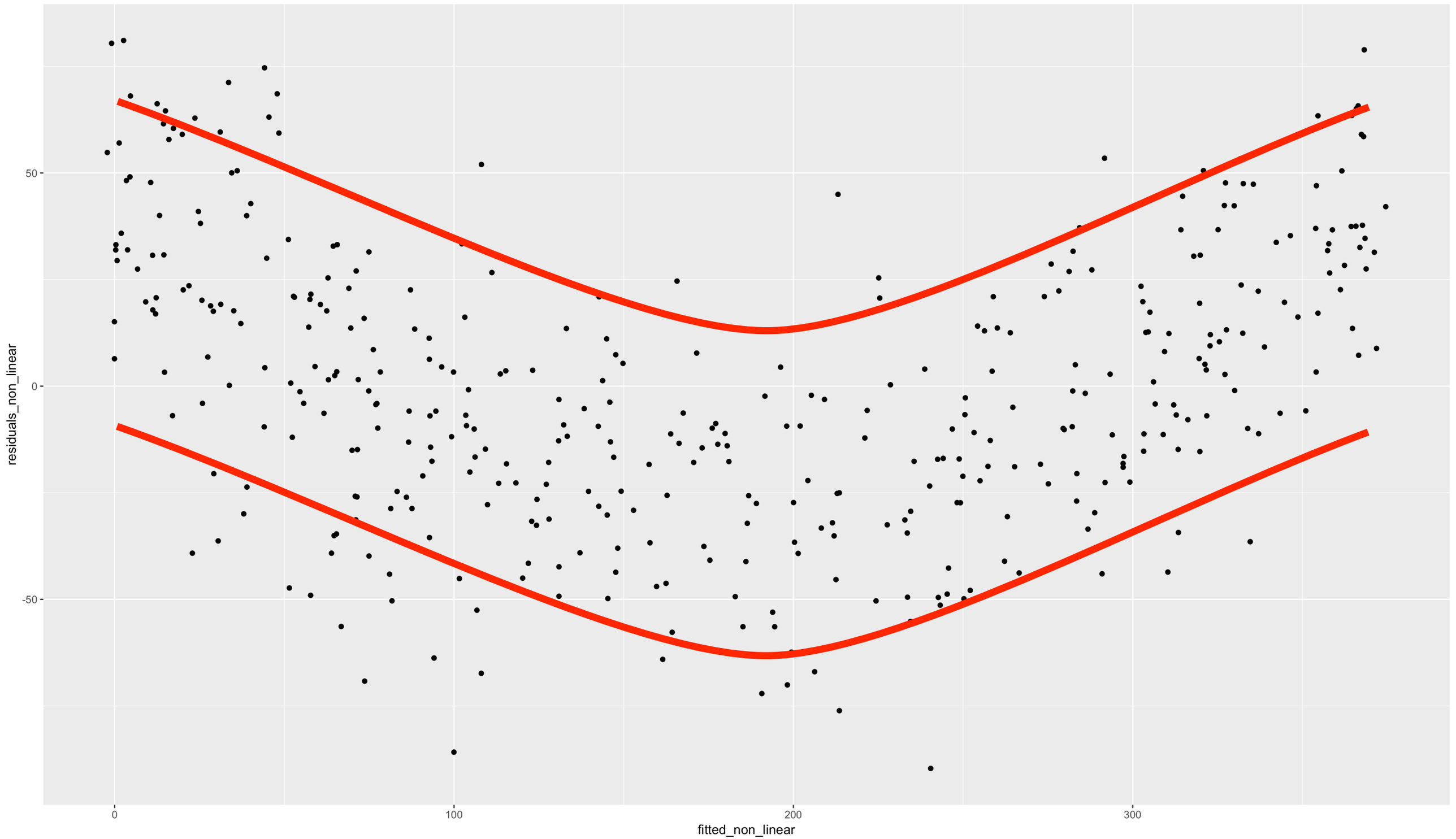
REGRESSION GRAPH



ERRORS NORMAL?



VARIANCE CONSTANT?



SUMMARY

Standardized residuals:

Examine normality

Residuals vs predicted y:

Homoskedasticity

Functional form

REMEDIES?

Non-normality:

More data

Heteroskedasticity:

Redefine X

Use another way to calculate standard errors

Functional form:

Experiment with other functional forms