Basic Linear Regression Analysis in R

Outline

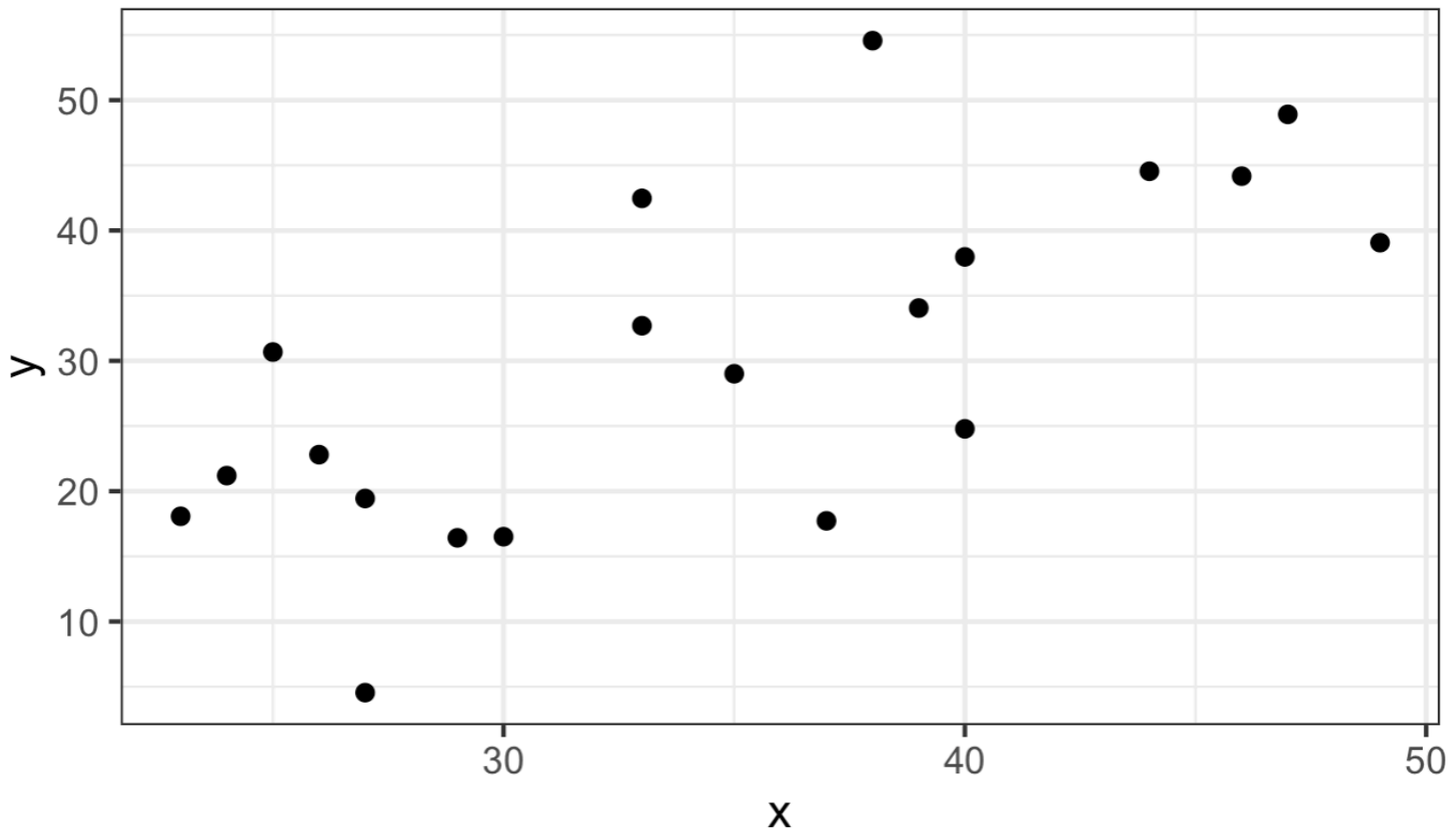
Motivation

Simple Linear Regression (Mathematical Viewpoint)

Statistical View of SLR

Other Considerations

Motivation

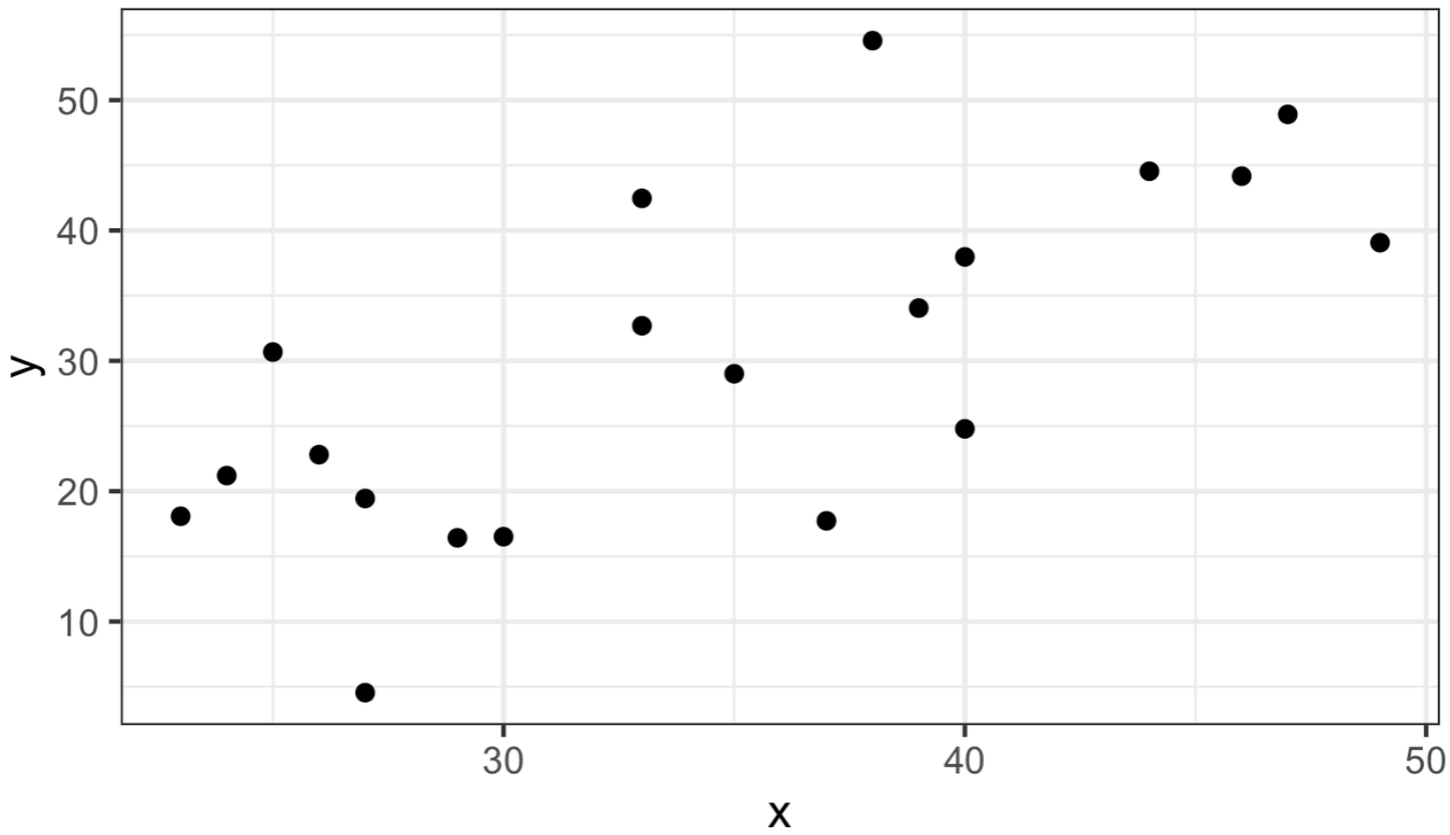


How can I model this relationship btw x and y?

Inference: How can I describe the relationship btw x and y?

Prediction: How can I accurately predict y given new values of x?

Motivation



Simplest Choice: A line!

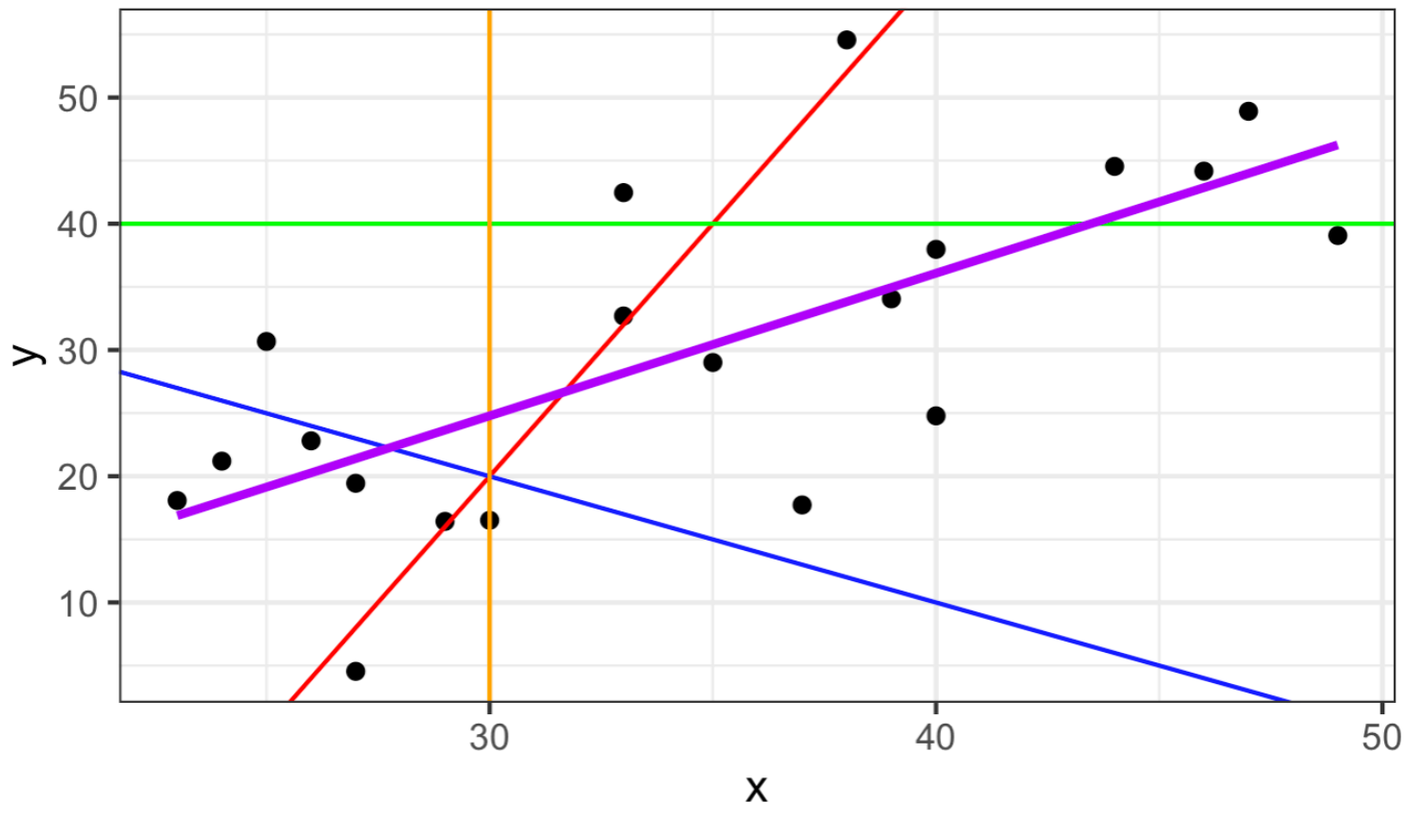
The process of using a line(plane) to model data is generally called linear regression

When there is only one predictor and one response, it is called simple linear regression.

Simple Linear Regression

Fitting a “line of best fit” to data is an optimization task that is completely mathematical in nature (calculus and linear algebra)

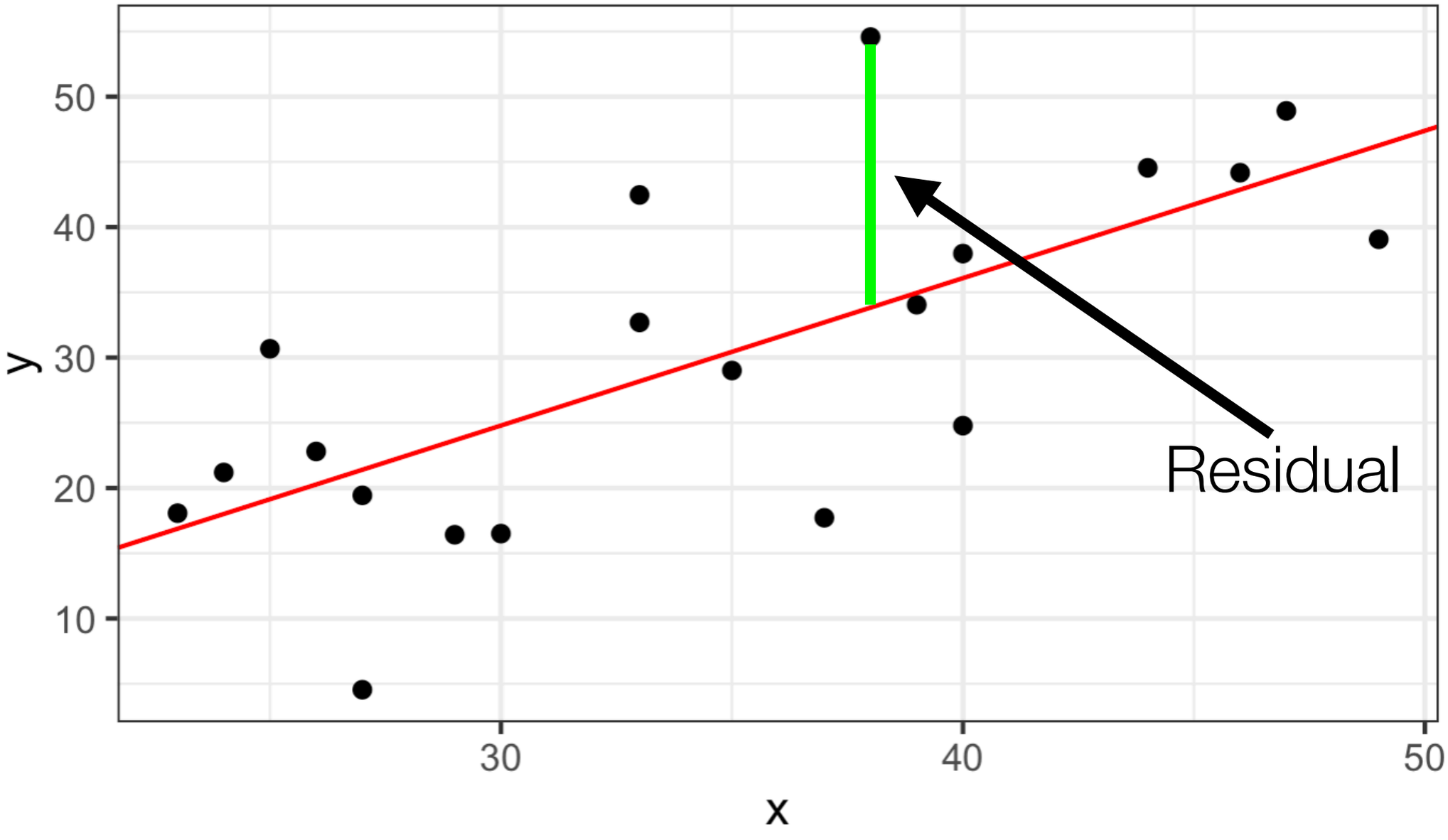
There are an infinite # of lines of the form $Y = \beta_0 + \beta_1 X$ that go through the data



Is it possible to say which is best? How do I define good/bad?

Simple Linear Regression

Residual (e_i) - The vertical difference between the actual and predicted y values for a particular x value



Least Squares Criterion

Let $e_i = y_i - \hat{y}_i$ represent the residual of the ith point where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the prediction for y based on the ith value of x.

Then the residual sum of squares (RSS) is defined as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

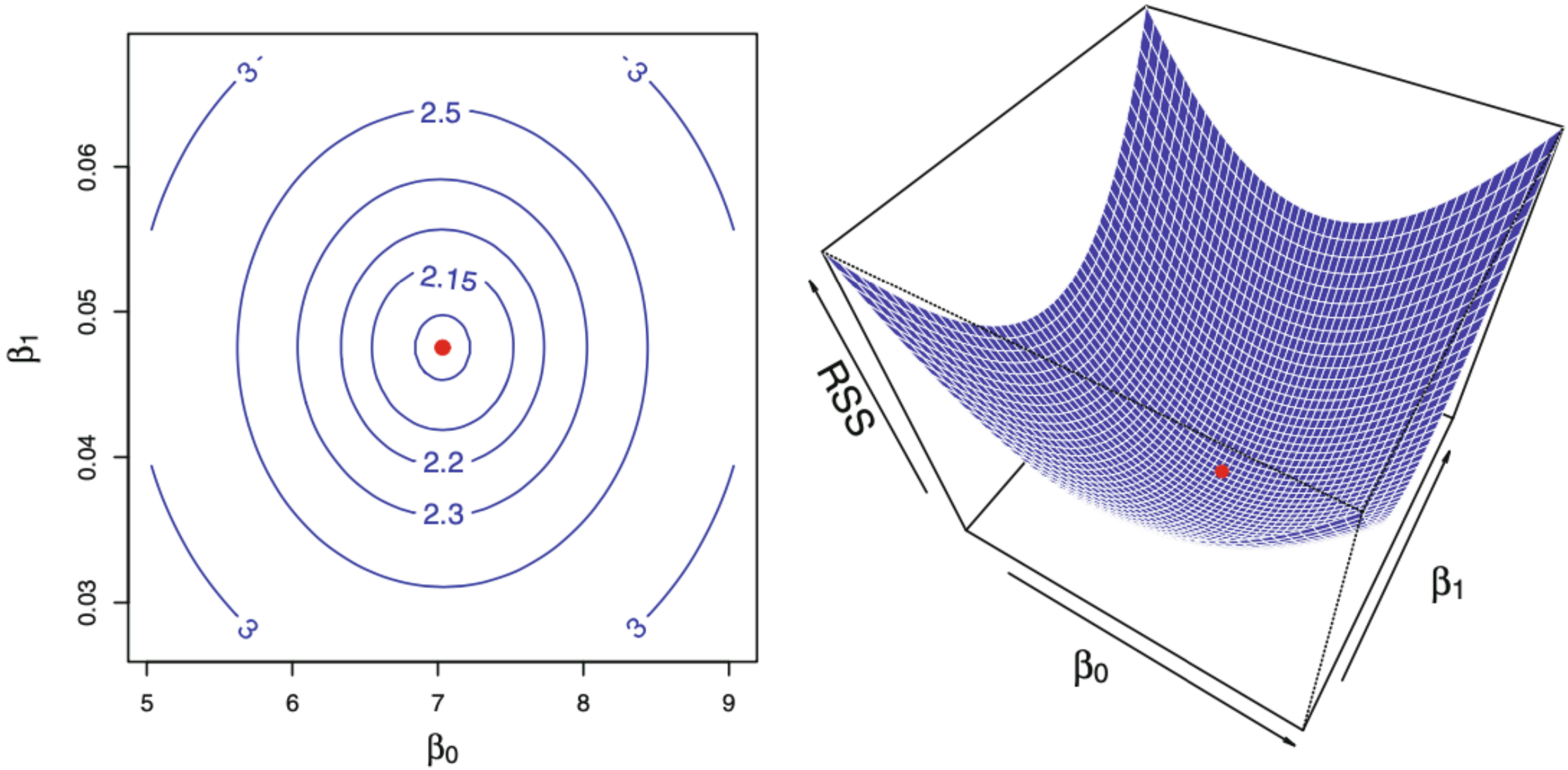
or equivalently as

$$\text{RSS} = [y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)]^2 + [y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_2)]^2 + \cdots + [y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)]^2$$

With some calculus we can find what values minimize RSS:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least Squares Criterion



ISLR

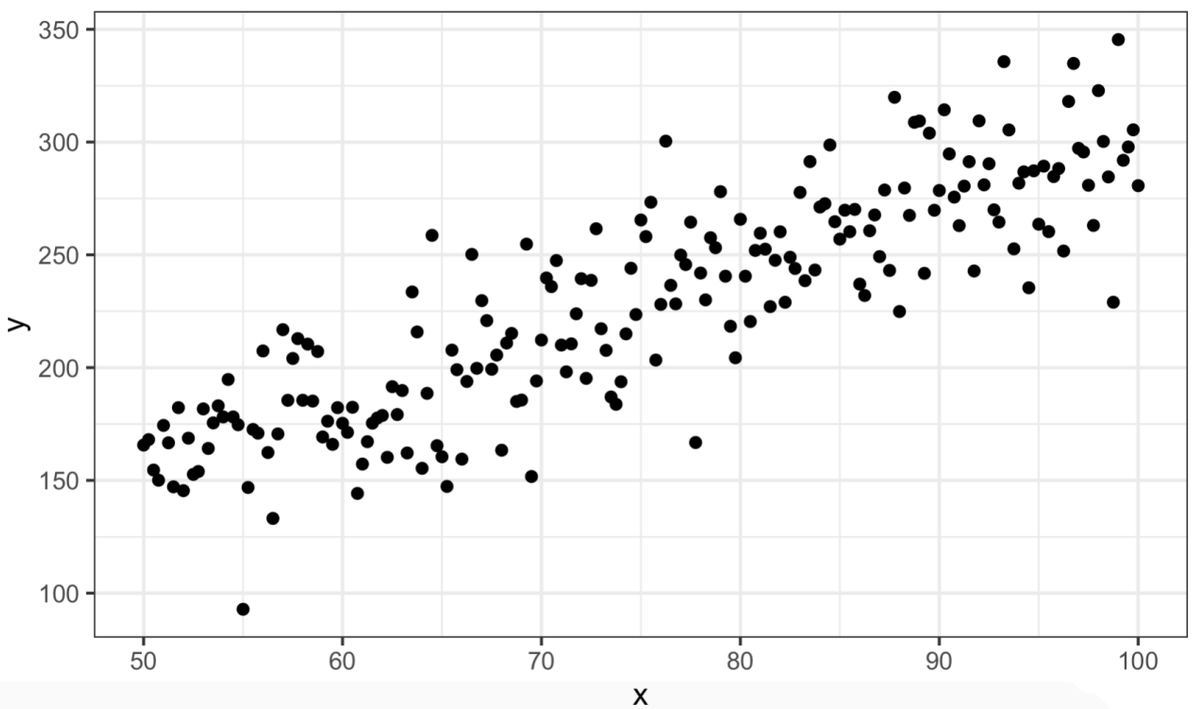
Linear Regression in R

Performing simple linear regression in R is super easy!

R uses a powerful formula syntax to describe the model.

Example: Simulated Data

```
data <- tibble(
    x = seq(50, 100, by = .25),
    y = 3*x + 5 + rnorm(length(x), 0, 25)
)
```




```
lm(y ~ x, data = data)
```

```
Call:

lm(formula = y ~ x, data = data)

Coefficients:

(Intercept)                x

      5.551            2.970
```

Linear Regression in R

Let's break down the syntax:

```
lm(y ~ x, data = data)
```

`lm:` Linear model - linear regression is a model that falls into a class of models where the predictors are a linear combination of one-another (linear models).

`y ~ x:` Formula syntax - the variable to the left of the `~` is the response variable and the collection of variables to the right are the predictors (the formulas can become quite complex!)

`data:` The data, of course!

SLR Statistical View

The only point at which statistics gets involved in linear regression is when the error term gets shows up!

General Form: $Y = f(\boldsymbol{X}) + \epsilon$

Simple Linear Regression: $Y = \beta_0 + \beta_1 X + \epsilon$

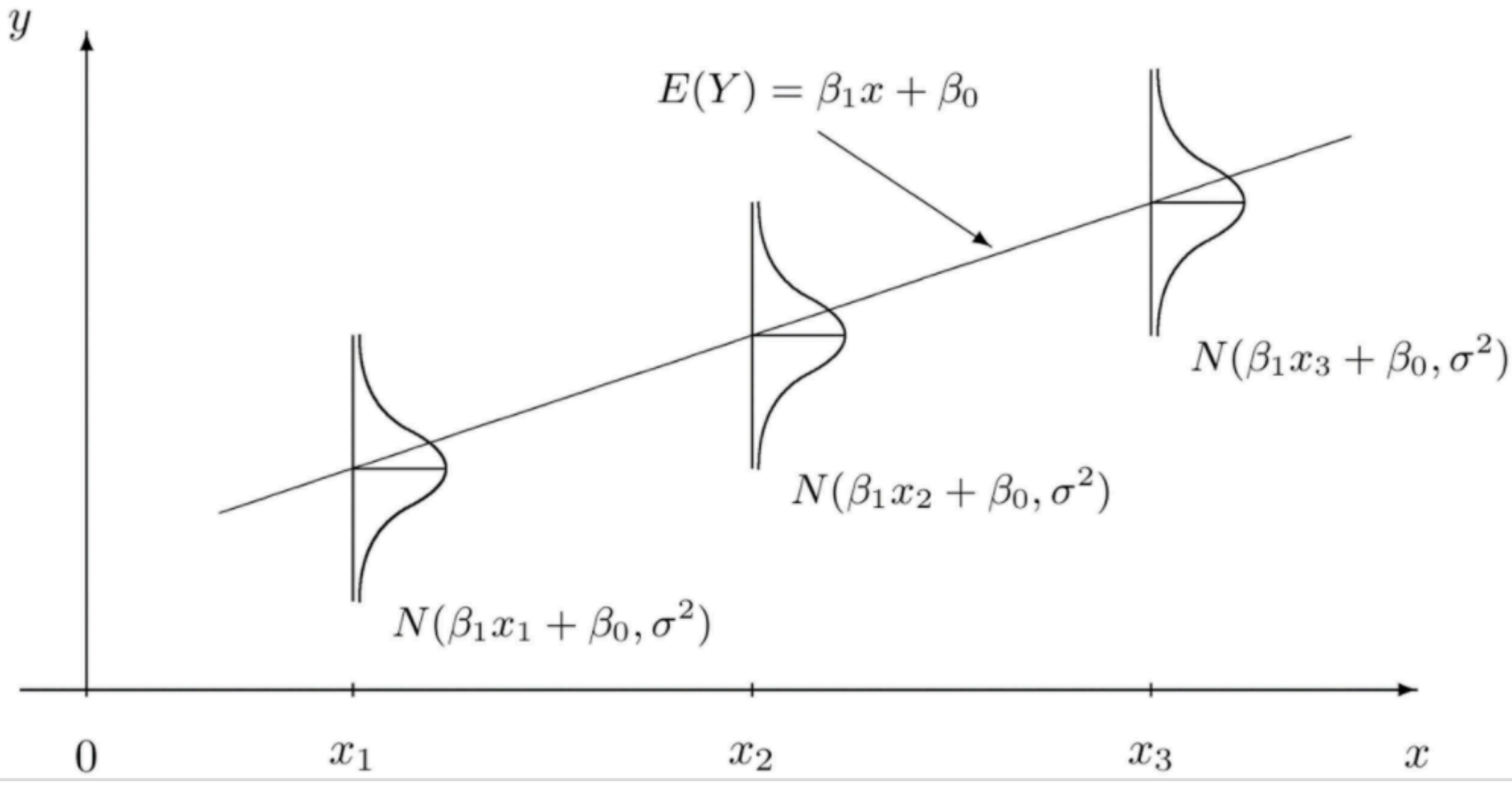
Theory of Linear Regression:

There is a true linear relationship between X and the mean of Y given X in the population: $\boldsymbol{\mu}_{Y|X} = \beta_0 + \beta_1 X$
(Crazy! I know!)

When we actually observe the data from the population, we observe the responses "with error". The error could occur for a number of reasons: $Y = \beta_0 + \beta_1 X + \epsilon$

We try to "recover" that true relationship with the least-squares line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

SLR Statistical View



Saylor Academy Intro Stats

SLR Statistical View

A number of assumptions have to be met in order to interpret statistical results from a linear regression:

1. The responses $y_1, y_2, \cdots, y_n$ come from a normal distribution, each with different means and a common variance (normality assumption)

2. The expected value (mean) of each y is a linear function of the predictors (linearity assumption)

3. The common variance is independent of the predictors (homogeneity of variance assumption)

4. The responses are independent (independence assumption)

SLR Inference in R

There's a lot more that we need to know (and hopefully do know from Stats II, but if not that's okay too!), but let's get to the fun stuff: Hypothesis tests, etc. for linear regression in R!

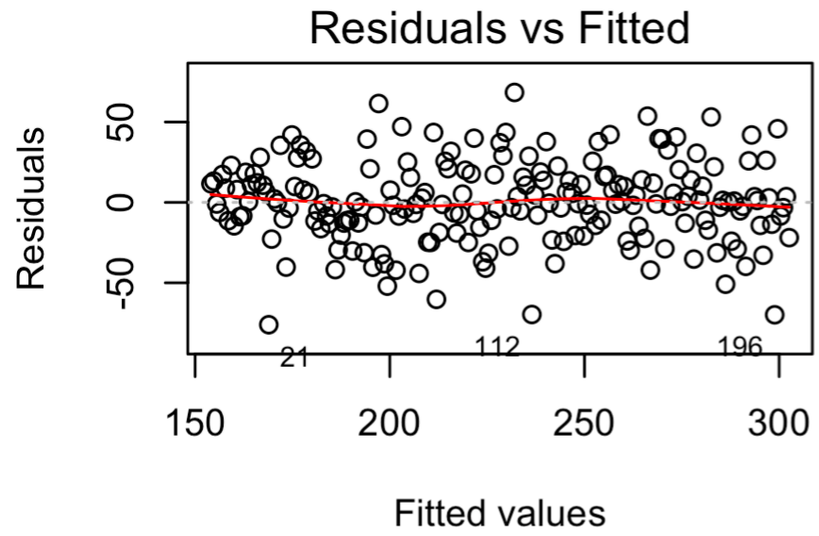
```
fit <- lm(y ~ x, data = data)

summary(fit)
```

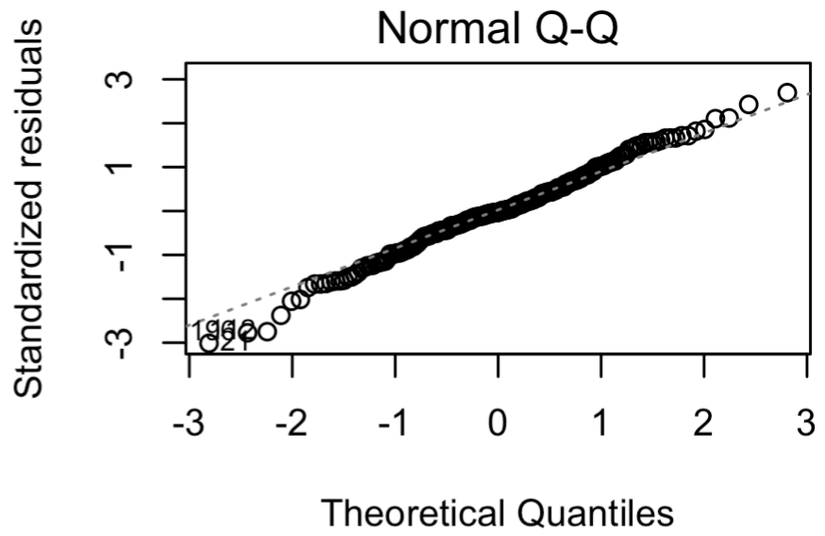
```
Call:

lm(formula = y ~ x, data = data)

Residuals:

    Min      1Q  Median      3Q     Max

-76.061 -14.284  -0.781  15.651  68.373

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   5.5511     9.4432   0.588    0.557

x             2.9703     0.1236  24.028   <2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.42 on 199 degrees of freedom

Multiple R-squared:  0.7437,    Adjusted R-squared:  0.7424

F-statistic: 577.3 on 1 and 199 DF,  p-value: < 2.2e-16
```

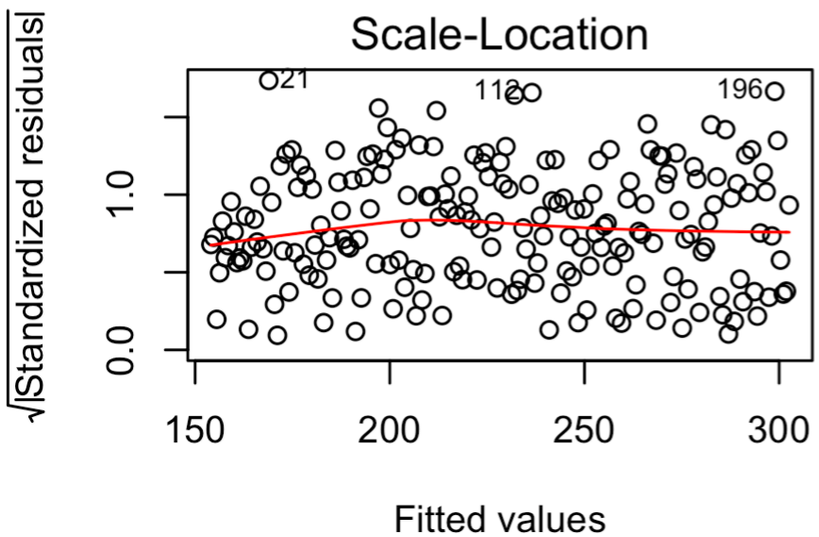
SLR Inference in R

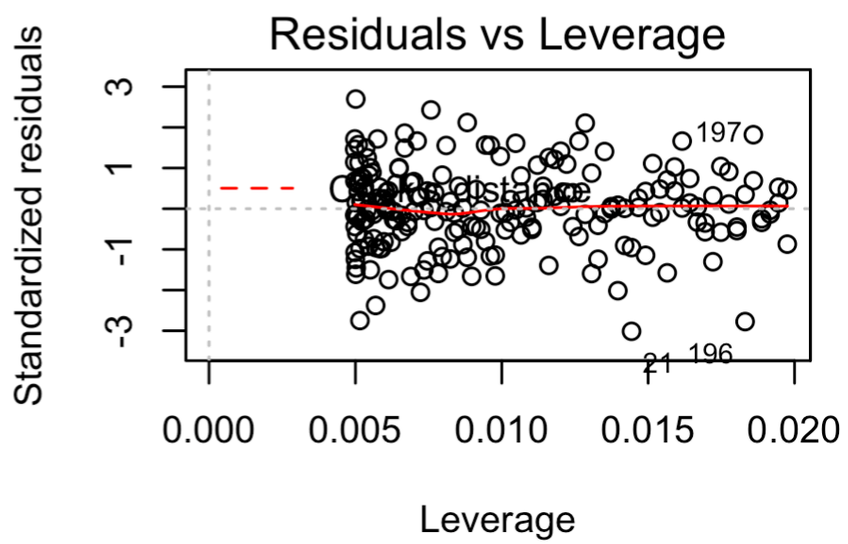
Graphical Checks for Assumptions:

```
par(mfrow = c(2,2))
plot(fit)
```









SLR Prediction in R

Prediction!

```
predict(fit, data.frame(x = c(62.3, 70.2)))
```

```
       1        2 

190.5983 214.0633 
```