Poli-301:

The Political Science Discipline

2

Correlations

Correlation

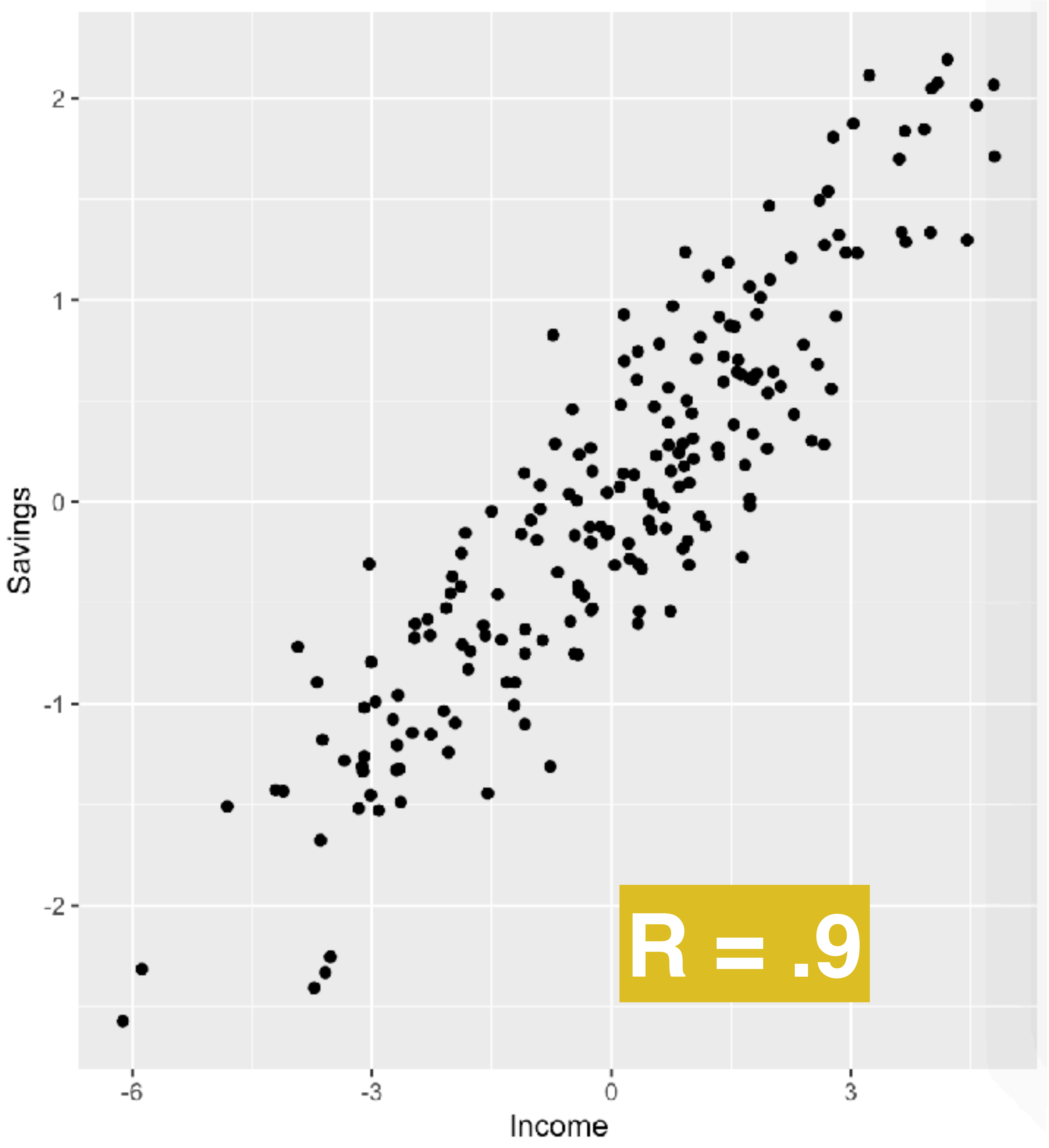
How closely two variables related

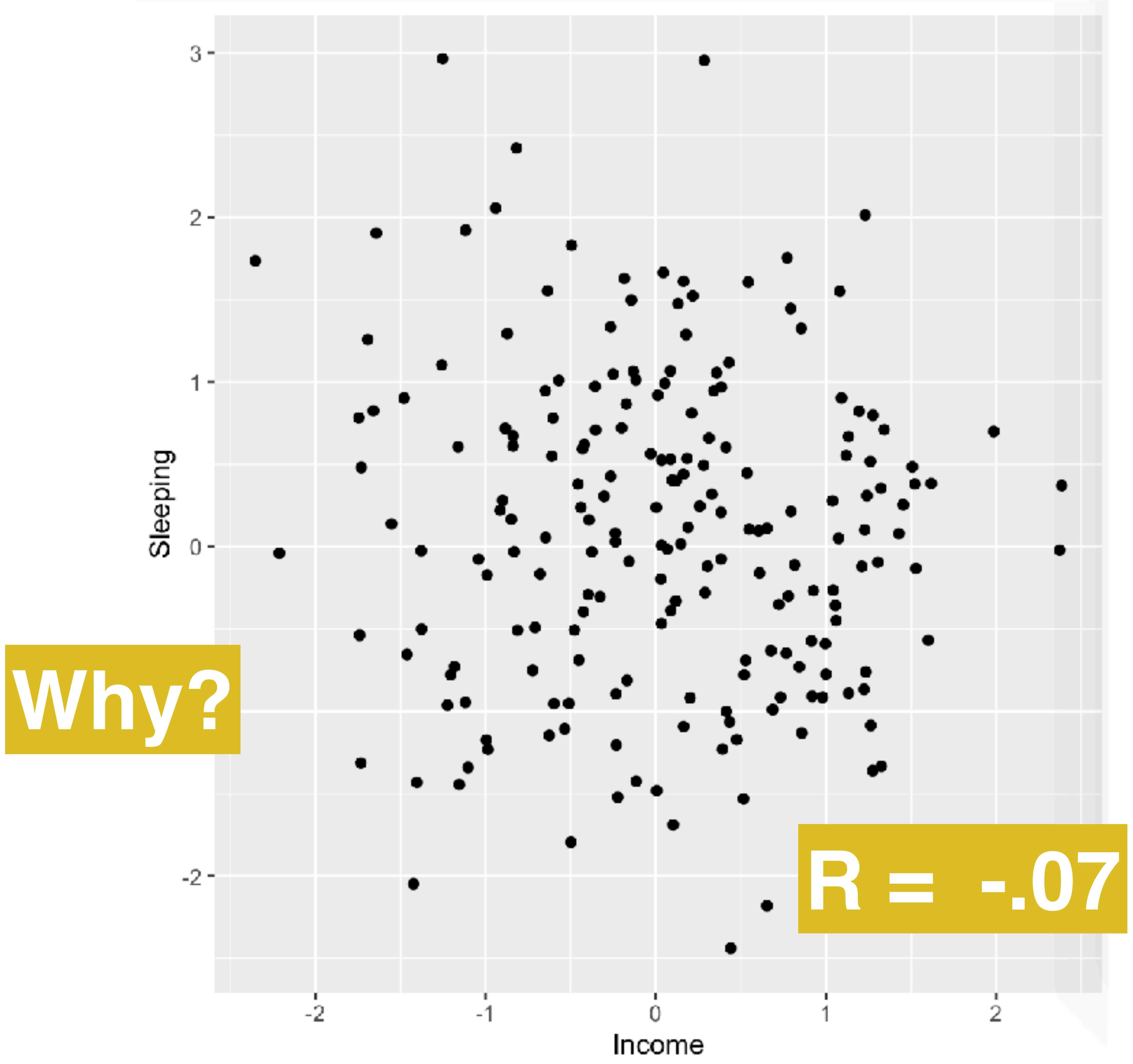
Direction of the relationship

-1 to 1

-1 and 1 = perfectly correlated
0 = perfectly uncorrelated

Relationships between variables





GENERAL GUIDELINES

0	No relationship
0.01–0.19	Little to no relationship
0.20–0.29	Weak relationship
0.30–0.39	Moderate relationship
0.40–0.69	Strong relationship
0.70–0.99	Very strong relationship
1	Perfect relationship

Can be positive or negative

Guess the Correlation

http://guessthecorrelation.com/

Math

https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/correlation-coefficient-r/v/calculating-correlation-coefficient-r

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Interpretation

As the value of X goes up,
Y tends to go up (or down)

A lot / a little / not at all

Relations between variables

What do we mean by strong "*relationship*"?

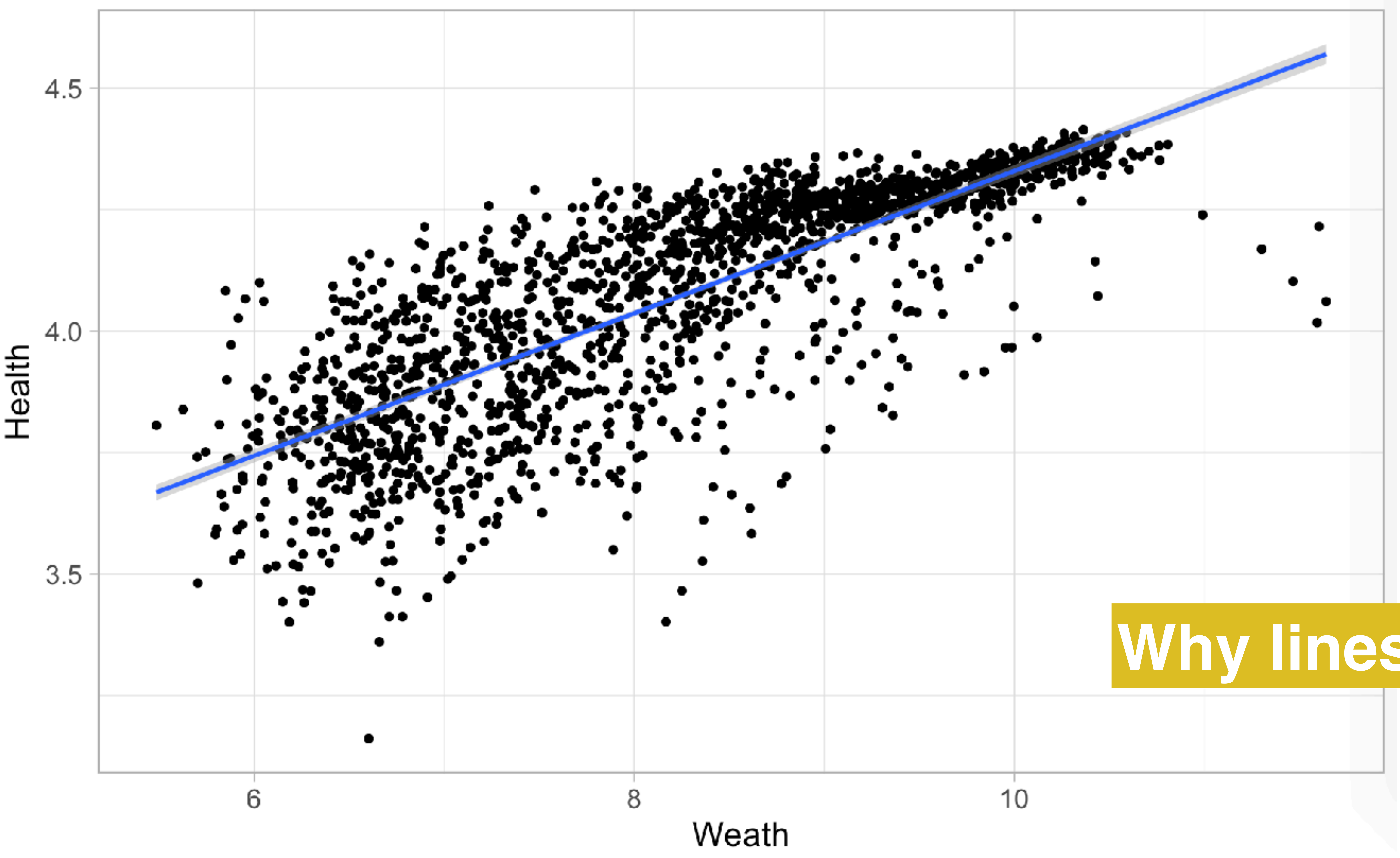
If X is high/low then
Y is high/low

Having information about X =
you know something about Y

3

Drawing Lines

Drawing Lines



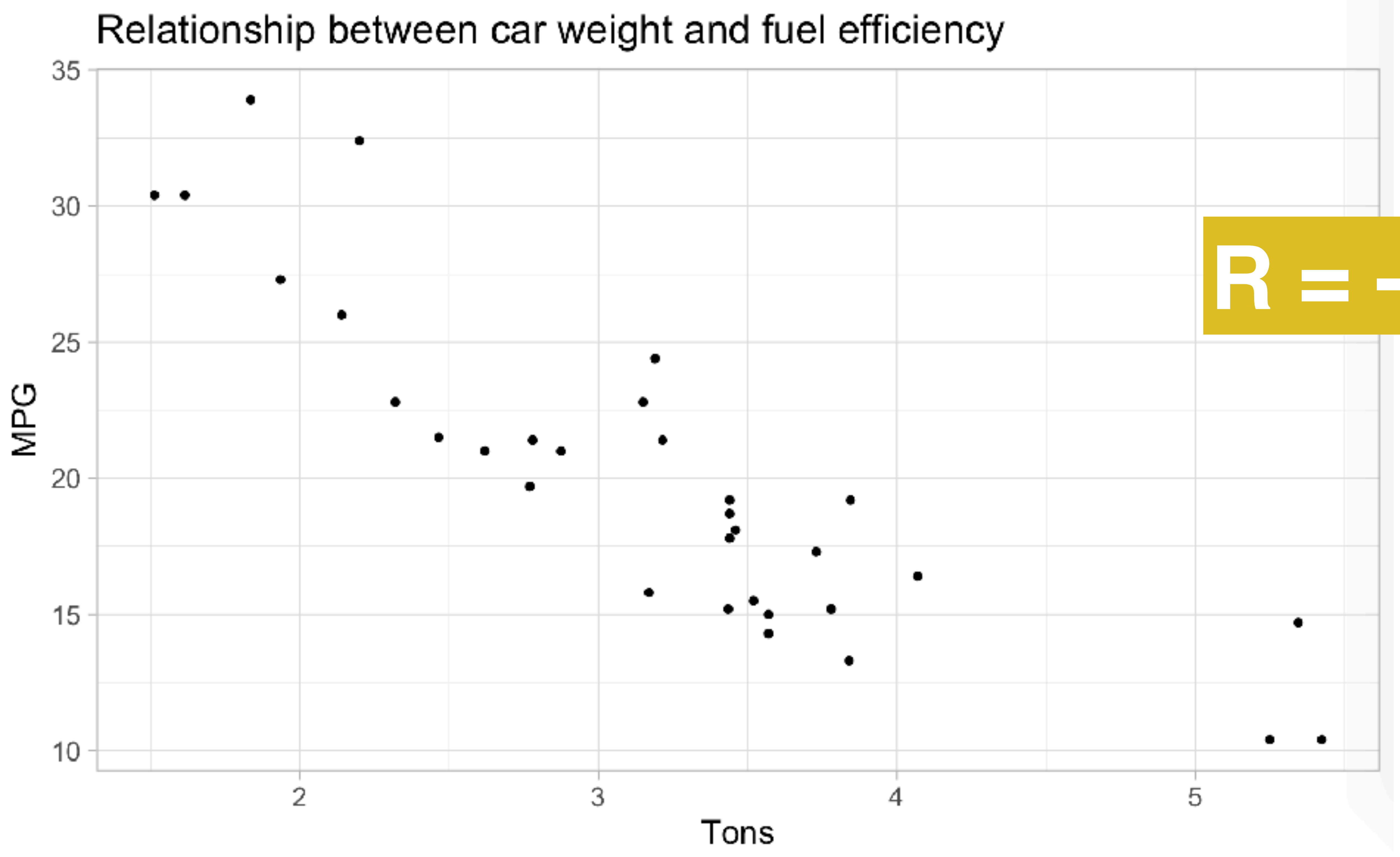
Why lines?

Why lines?

Correlation just gives us
direction and strength of relationship

But what if we wanted to know
what level of health we should expect
given a level of wealth?

Cars



R = -.87

If we took a ton off a car, how does MPG change?

What MPG should we expect given:
a weight of 3 tons,
A weight of 6 tons?

Alternative approaches

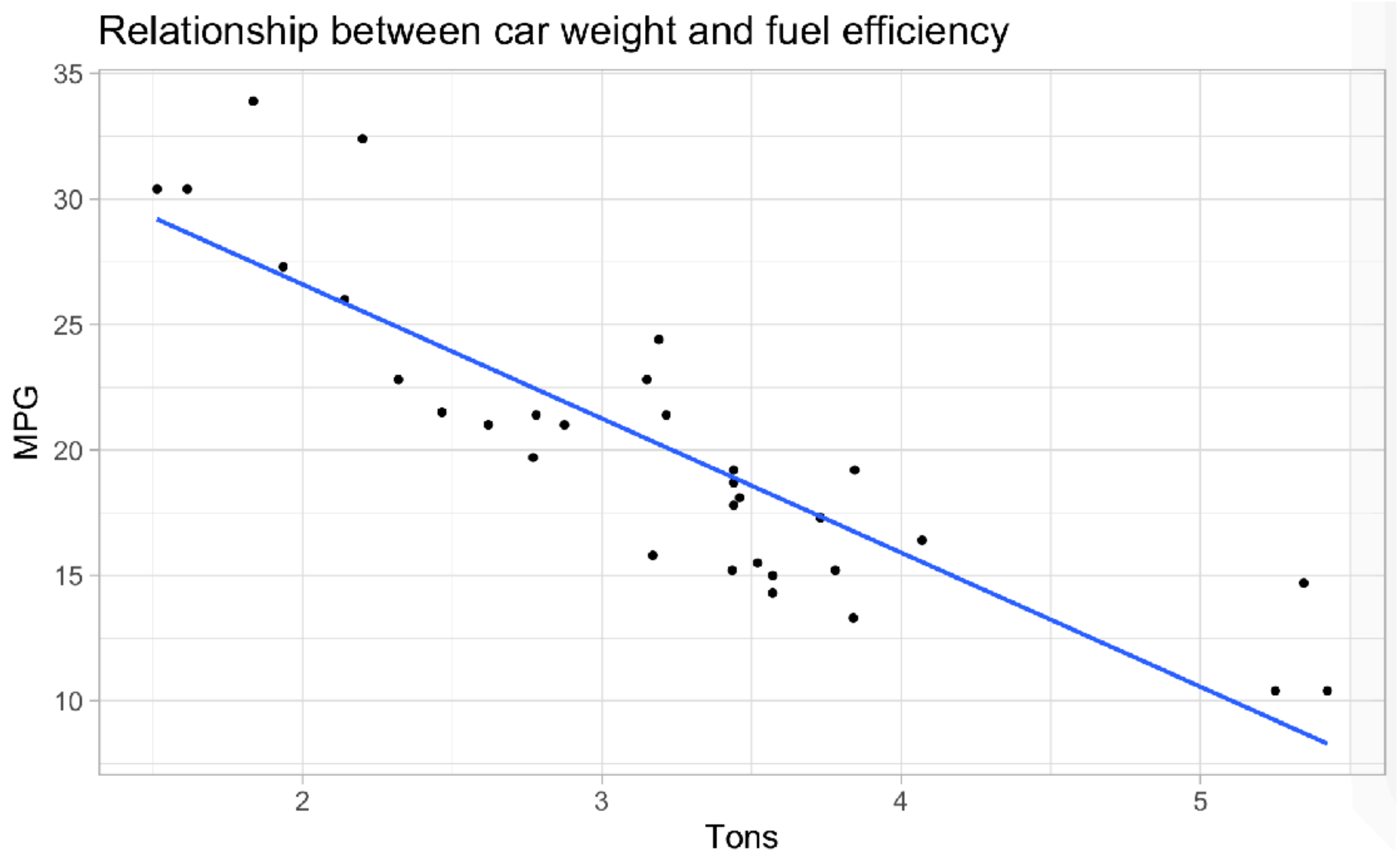
```r
# average MPG per weight
mtcars %>%
  group_by(wt) %>%
  summarise(avg_mpg = mean(mpg))
```

```
   wt avg_mpg
<dbl>   <dbl>
 1.51    30.4
 1.62    30.4
 1.84    33.9
 1.94    27.3
 2.14    26
 2.2     32.4
 2.32    22.8
 2.46    21.5
 2.62    21
 2.77    19.7
… with 19 more rows
```

```r
# average MPG per weight categories
mtcars %>%
  mutate(wt_cat = cut_number(wt, n = 3)) %>%
  group_by(wt_cat) %>%
  summarise(avg_mpg = mean(mpg))
```

```
wt_cat        avg_mpg
<fct>           <dbl>
[1.51,2.81]      26.1
(2.81,3.5]       19.4
(3.5,5.42]       14.7
```

Strengths vs. weaknesses?



Guess MPG even where no data on weight

Slope of line gives me general rate of change

Speaking the language

Y	~	X
Outcome variable		Explanatory variable
Response variable		Predictor variable
Dependent variable		Independent variable
What you want to explain or predict		What you use to explain changes in Y

Identify the parts

A car company determining the effect of weight on fuel efficiency

Weather channel using changes
in temperature, pressure, humidity, etc.,
to predict rain or no rain for tomorrow

Do students who get tutoring
tend to get higher grades?

Amazon using your browsing history,
past purchases, demographics, etc., to suggest purchases

Steps

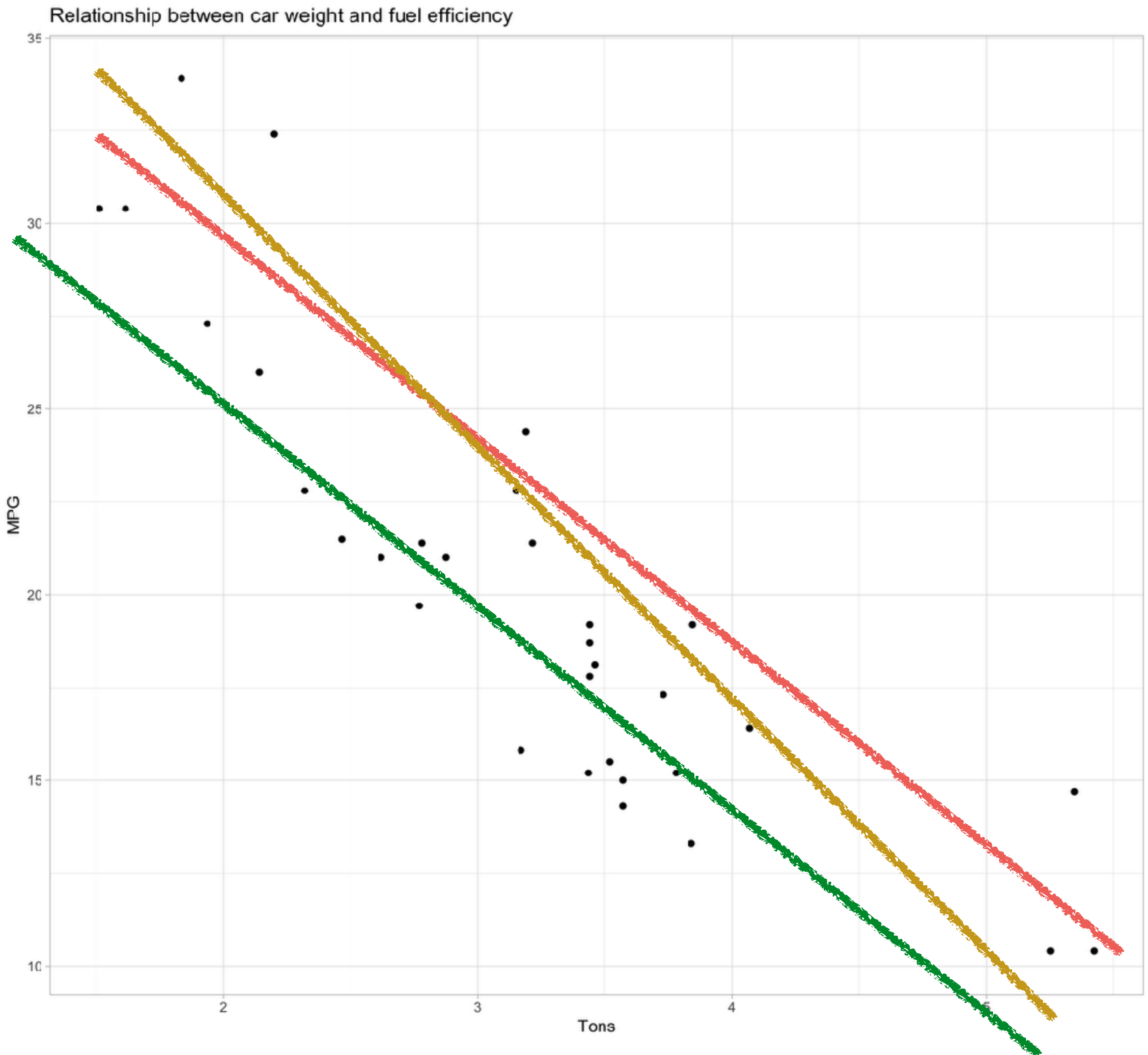
Plot X and Y

Draw line to approximate relationship

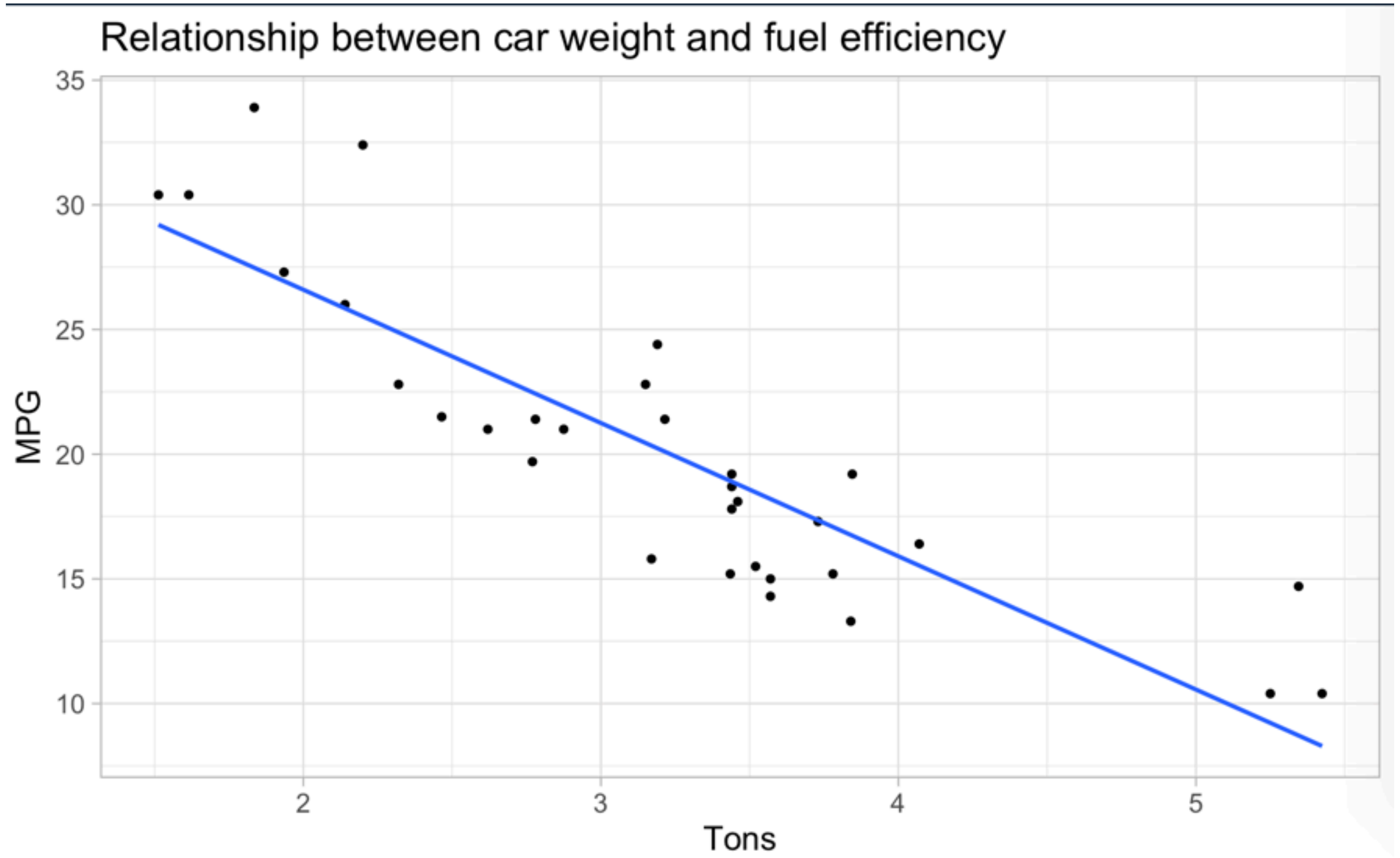
Find “math-y” parts of line

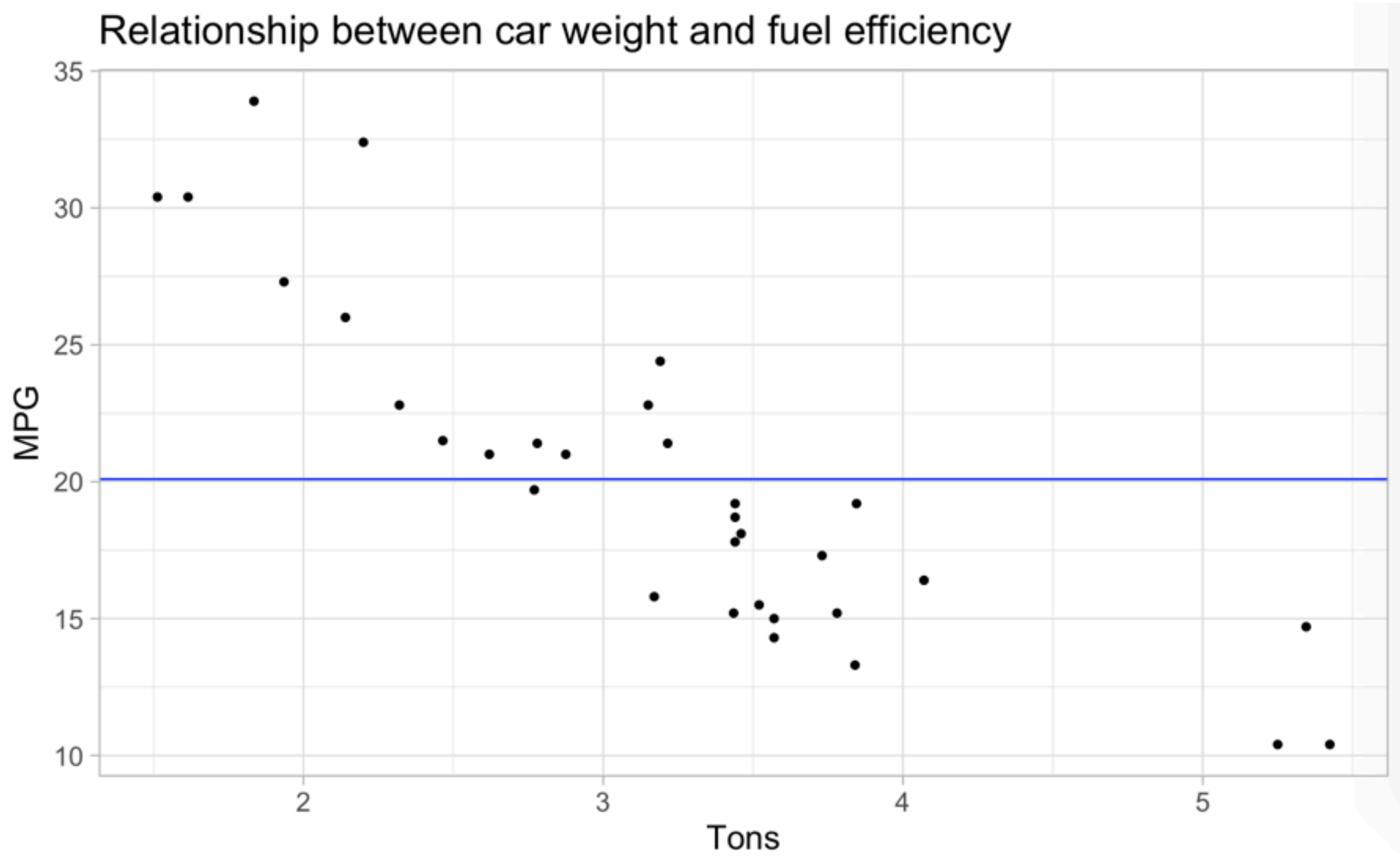
Interpret the math

Which line to draw?

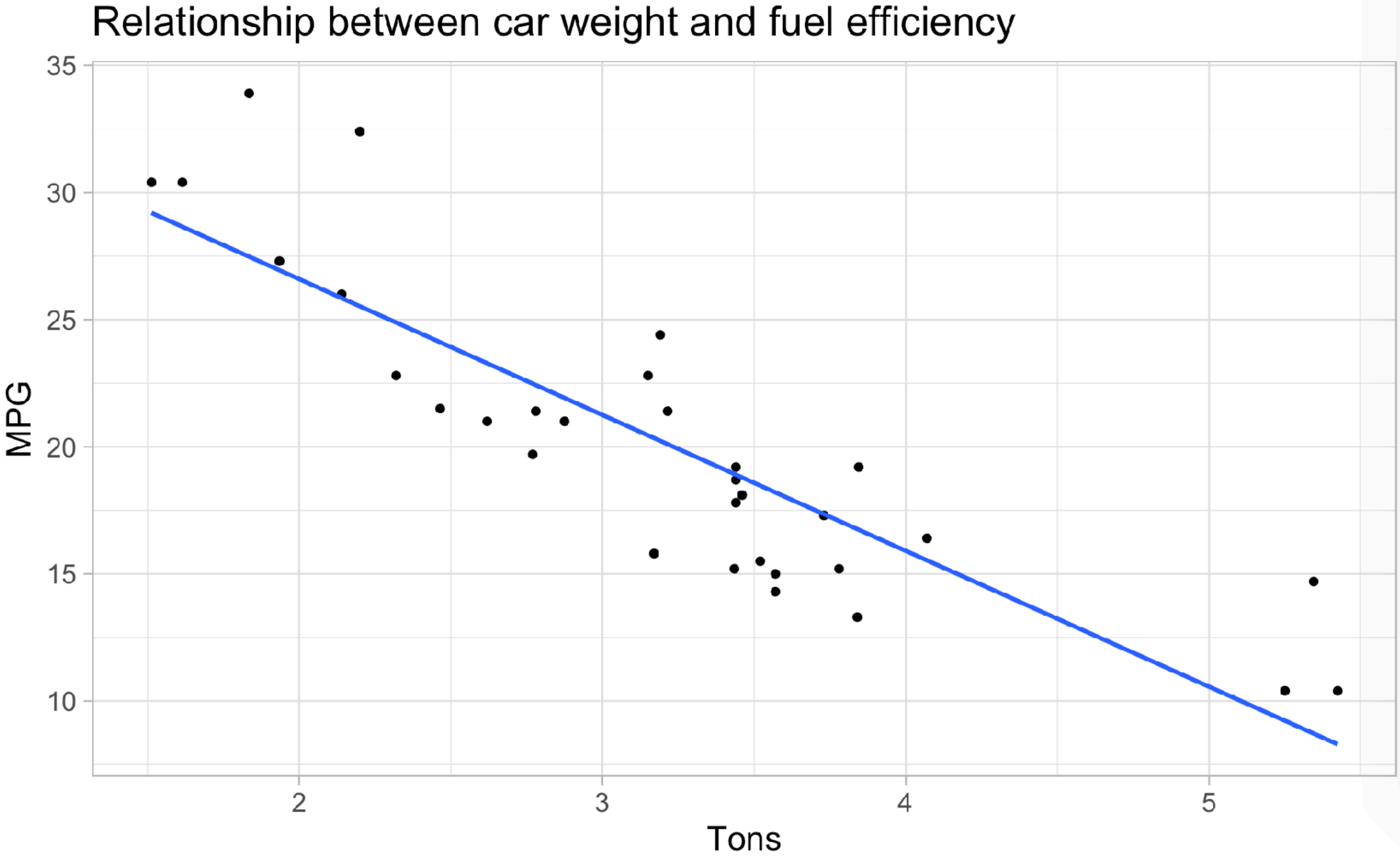


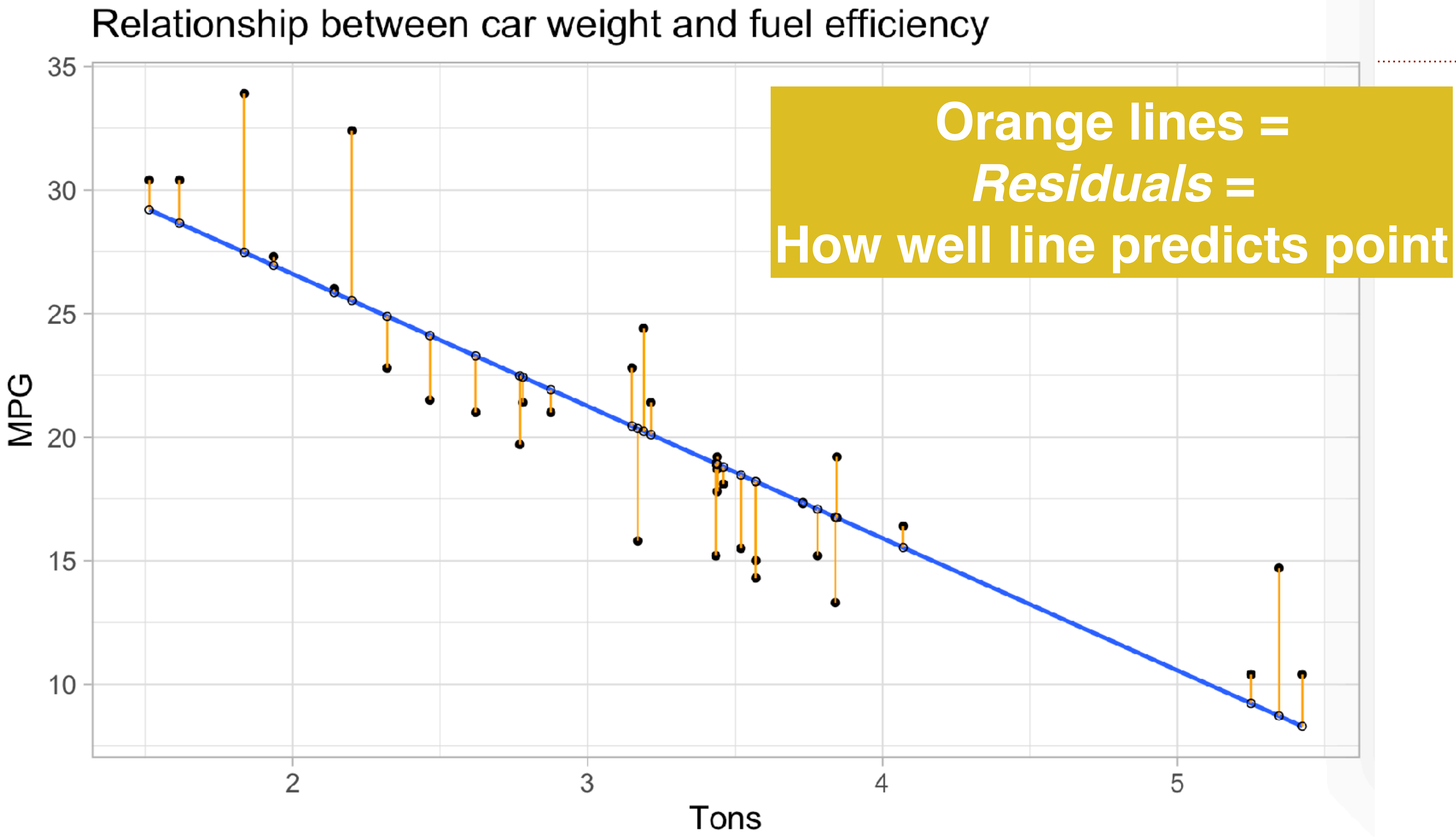
Compare two lines

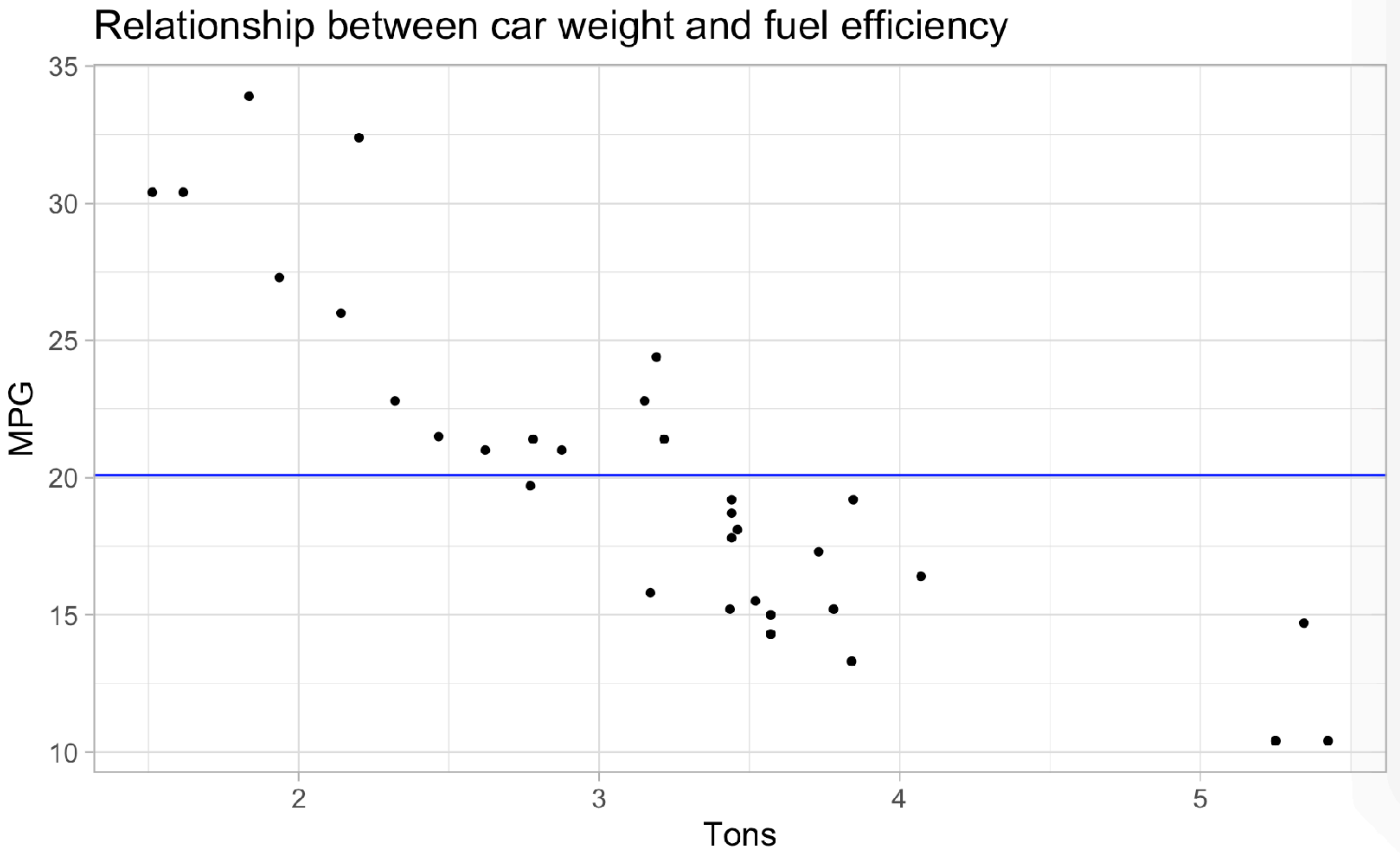


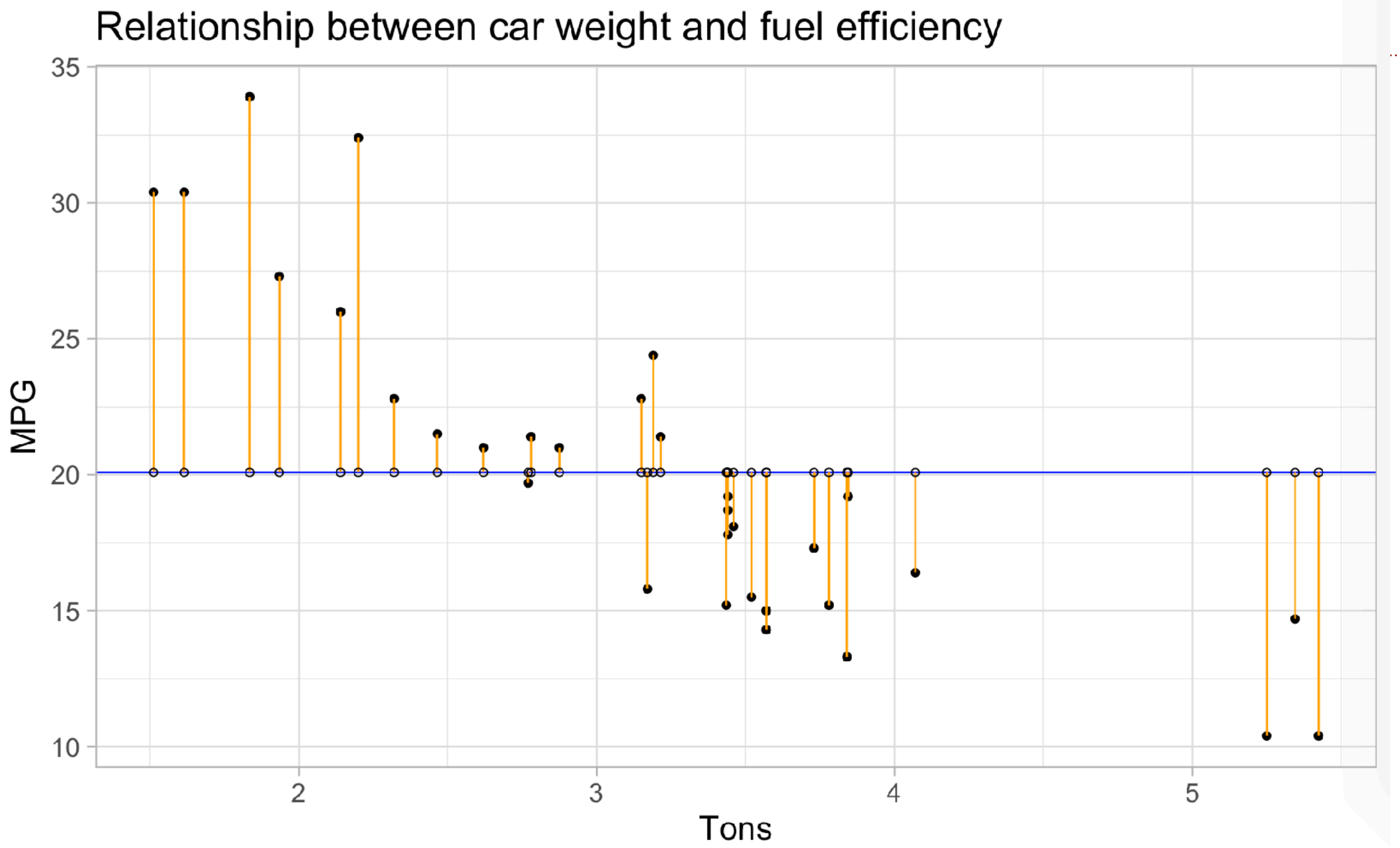


Which line fits better? Why?

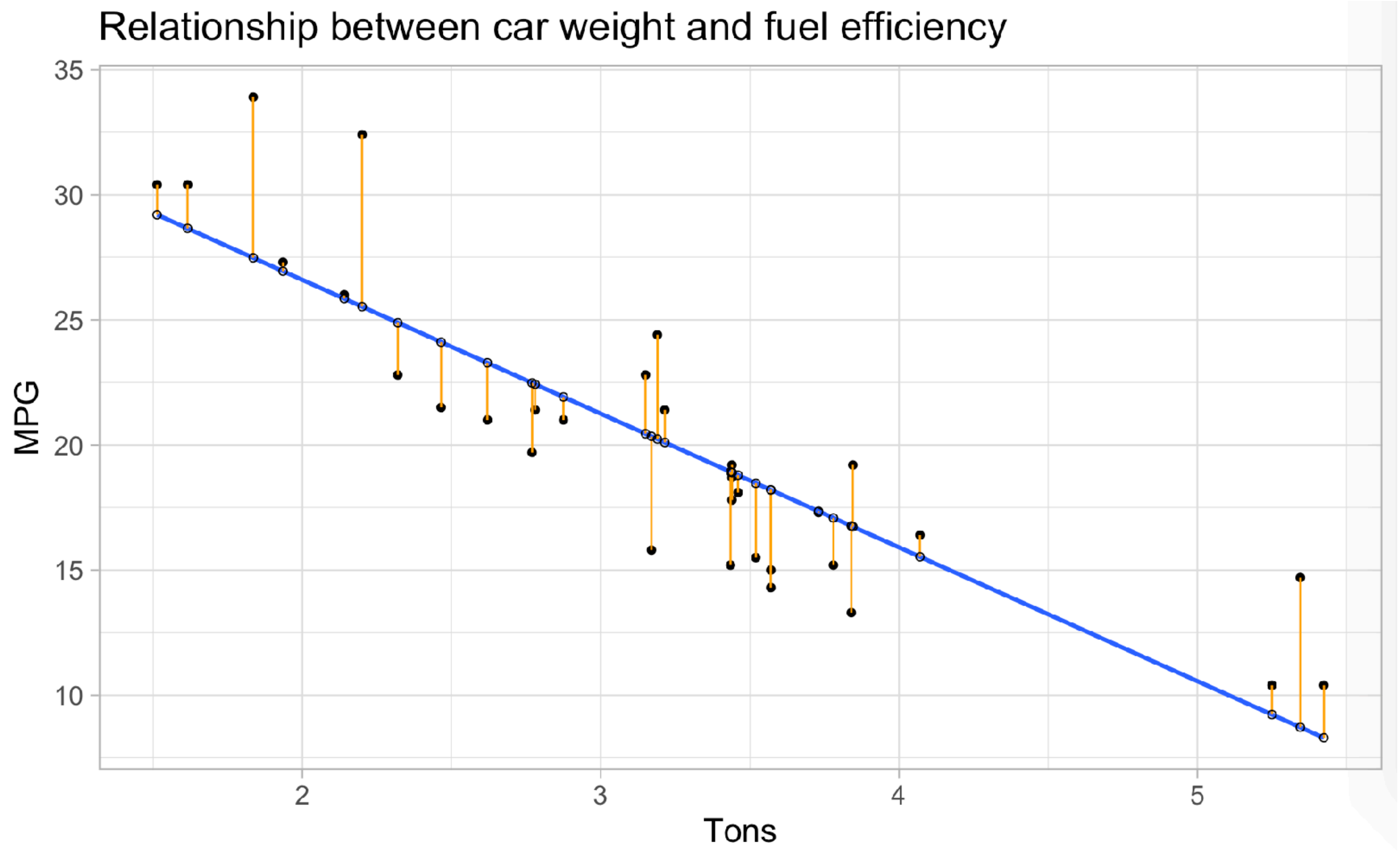

Relationship between car weight and fuel efficiency
35
30
25
20
15
10
MPG
2
3
4
5
Tons

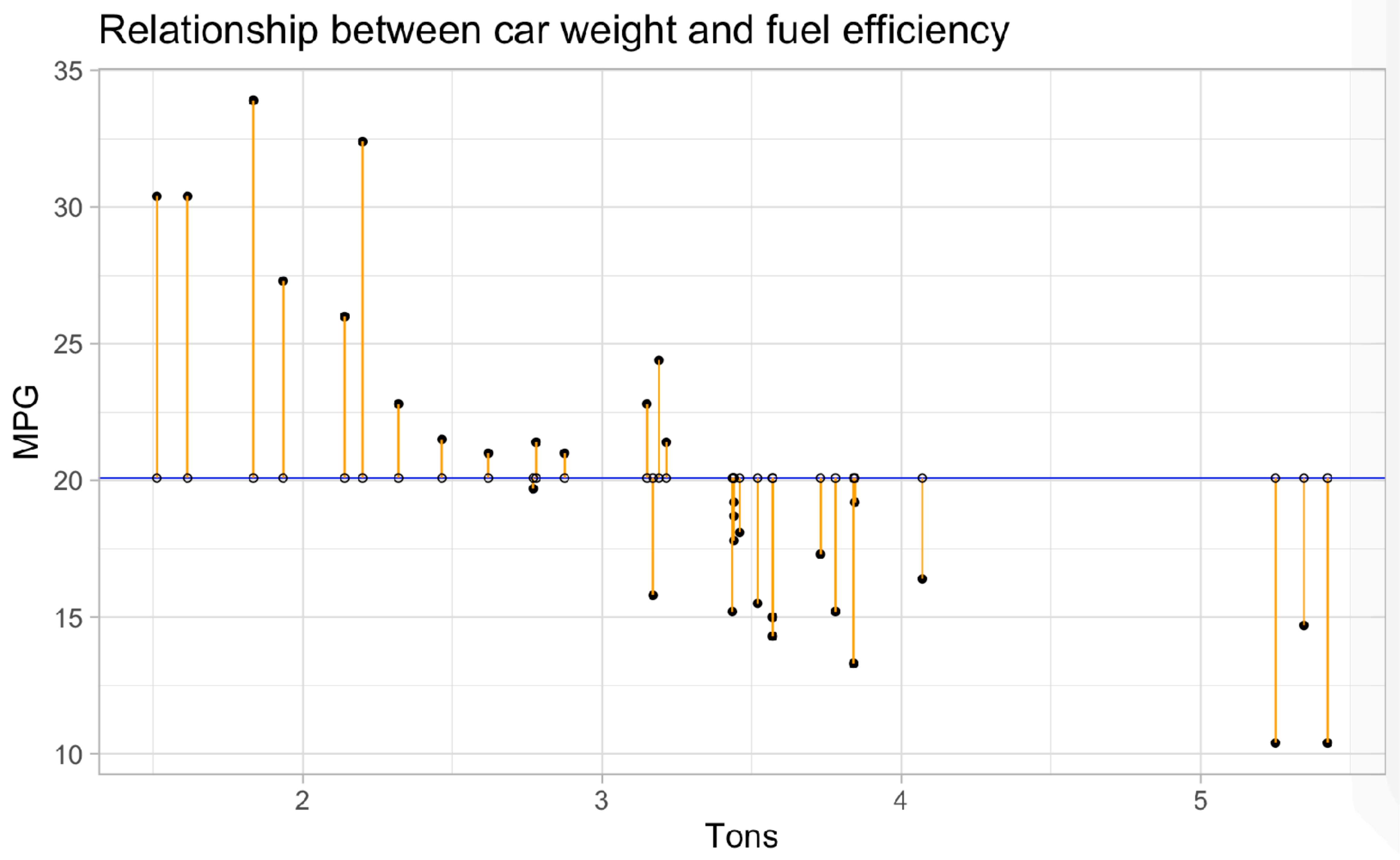

Relationship between car weight and fuel efficiency
Orange lines =
Residuals =
How well line predicts point
35
30
25
20
15
10
MPG
2
3
4
5
Tons


Relationship between car weight and fuel efficiency
MPG
Tons
35
30
25
20
15
10
2
3
4
5


Relationship between car weight and fuel efficiency
MPG
Tons
35
30
25
20
15
10
2
3
4
5

Which is better?





SSR = 278

For each potential line:
Square distances from line,
Add them all up
Pick line with lowest sum

SSR = 1126

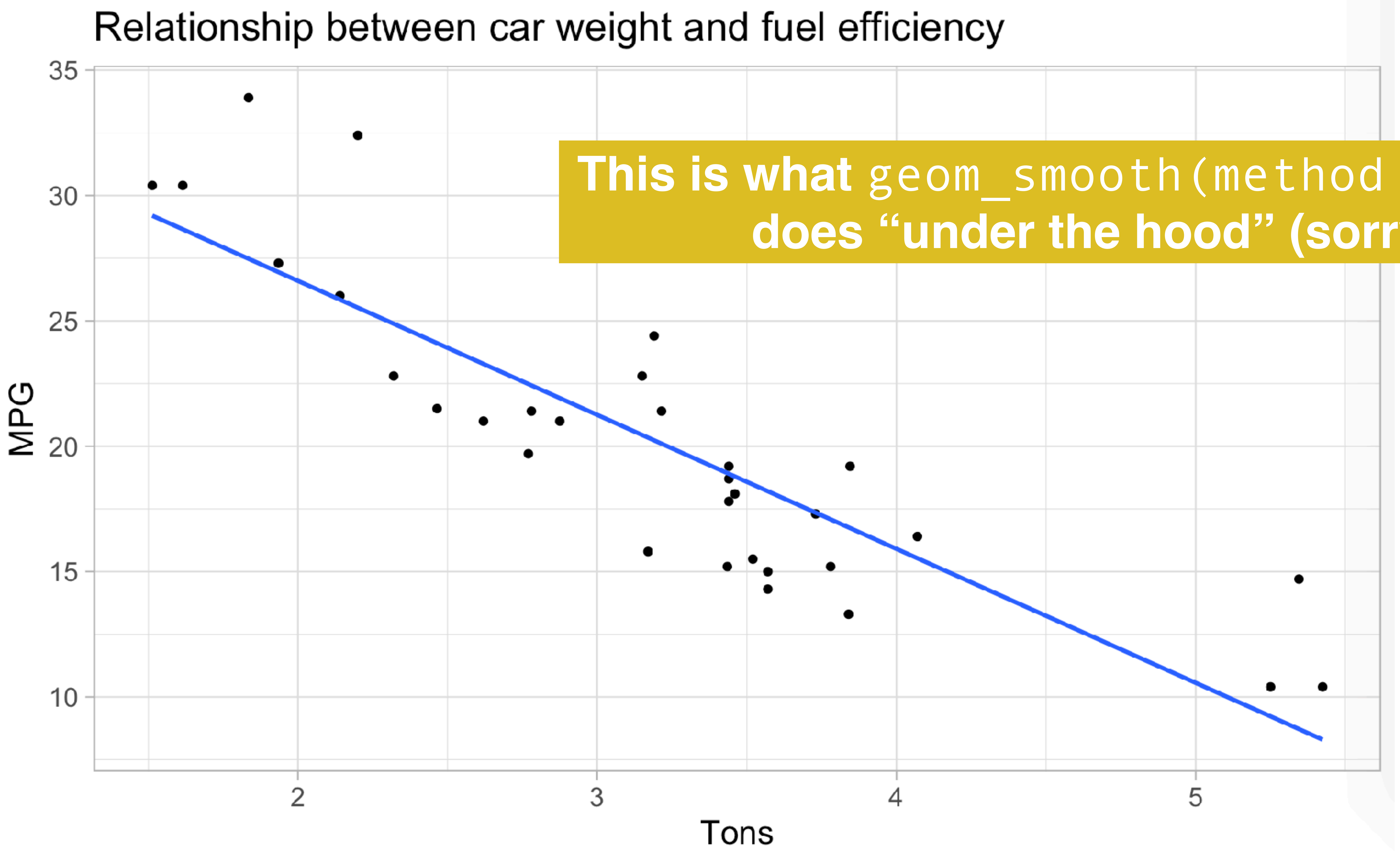
The best line

Ordinary Least Squares (OLS) is one way of getting "best" line

Minimize sum of the squares of the differences between the observed and predicted values

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

OLS



This is what `geom_smooth(method = “lm”)` does “under the hood” (sorry)

Getting the line

$Y = mx + b$

Y = a number

x = a number

m = slope

$\frac{rise}{run}$

b = y-intercept

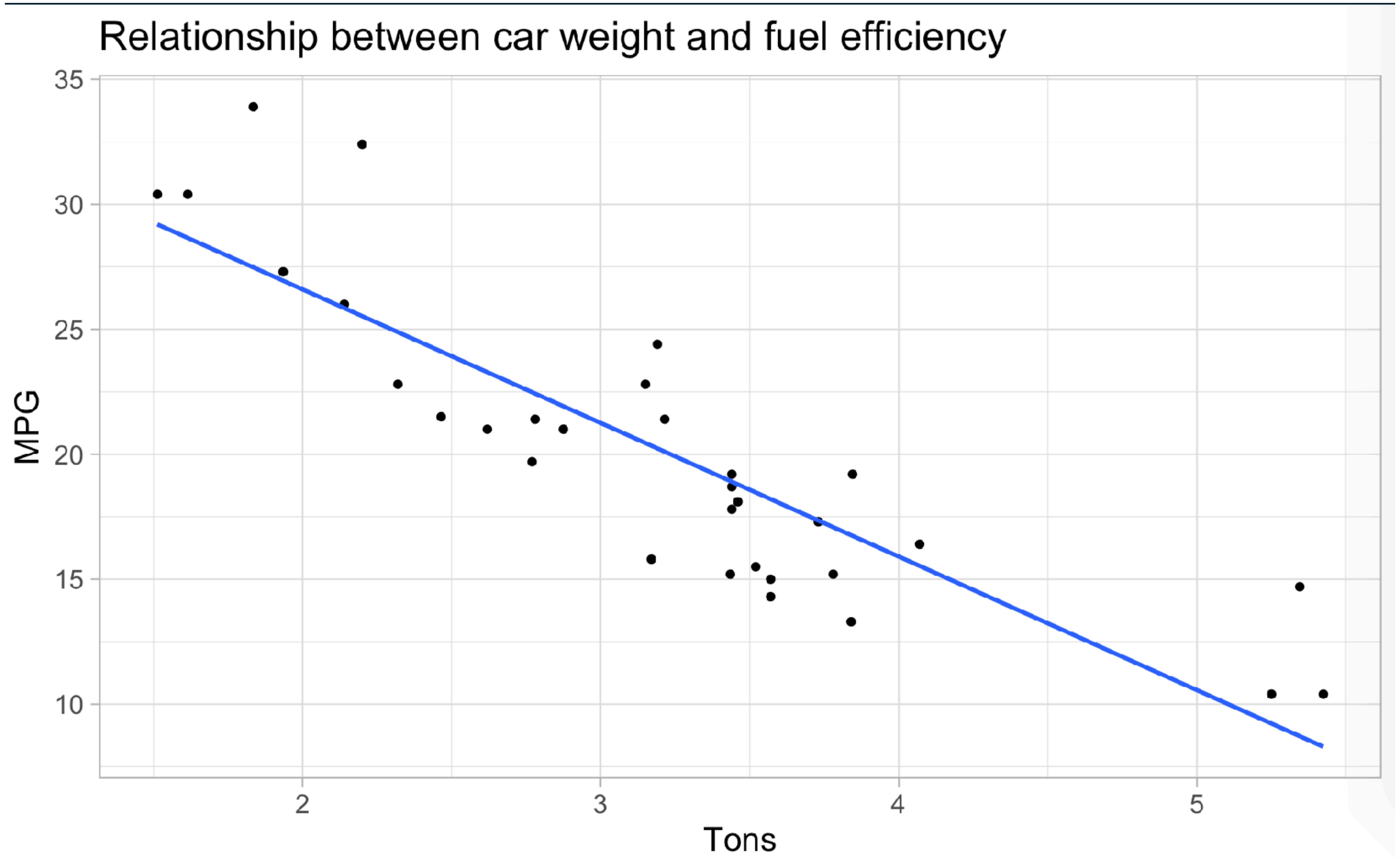
From 6th grade algebra to Statistics

Y = mx + b	$\hat{y} = \beta_0 + \beta_1 x_1 + \epsilon$	
Y	$\hat{y}$	Outcome variable
x	x_1	Explanatory variable
b	β_0	Y intercept
m	β_1	Slope
	ϵ	Error (residual)

Modeling car weight and fuel efficiency

$$\hat{y} = \beta_0 + \beta_1 x_1 + \epsilon$$

$$\hat{mpg} = \beta_0 + \beta_1 weight + \epsilon$$

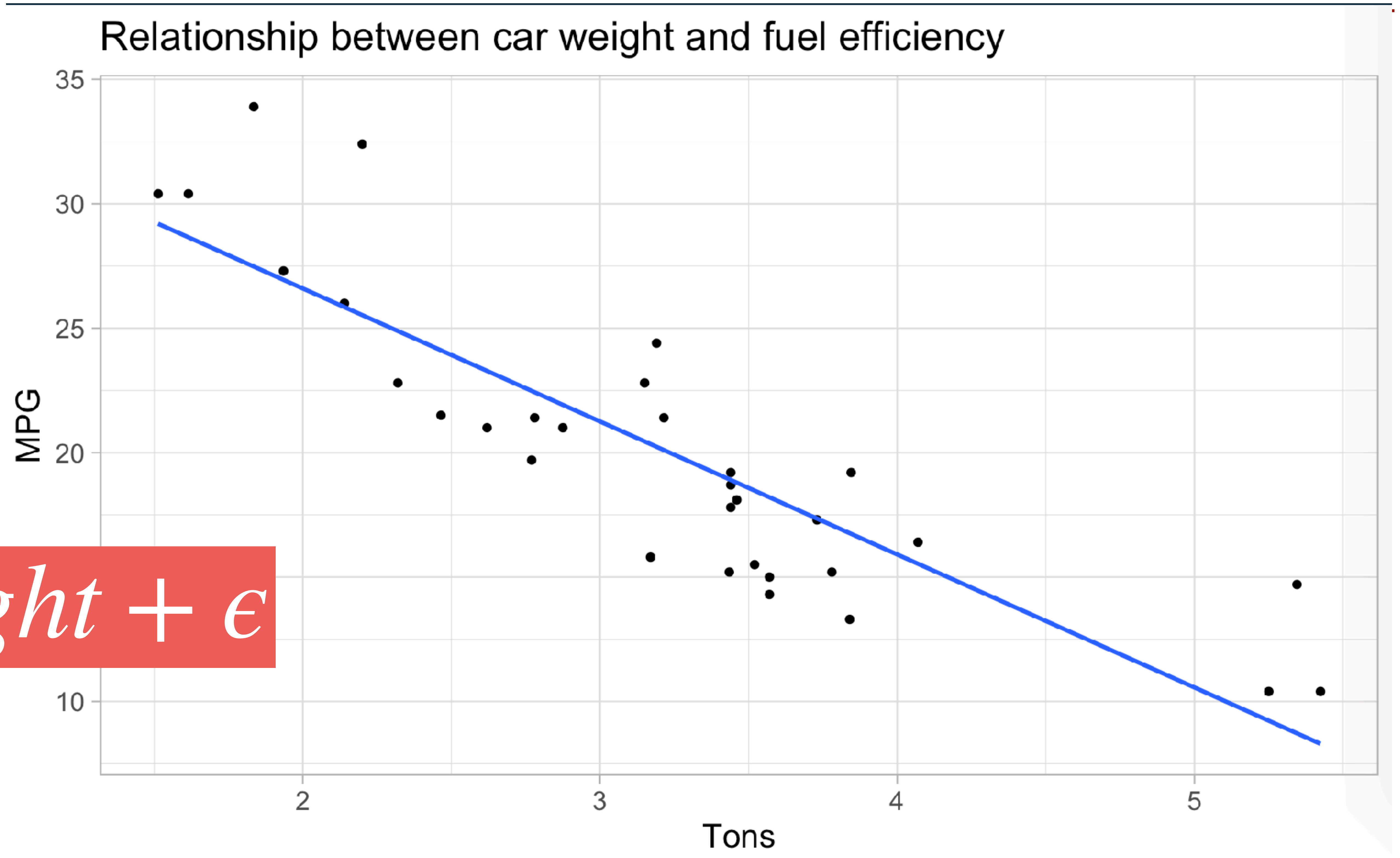


Modeling car weight and fuel efficiency

```
mpg_model = lm(mpg ~ wt, data = mtcars)

mpg_model %>%
  get_regression_table()
```

```
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    37.3      1.88      19.9        0    33.4     41.1 
2 wt           -5.34     0.559     -9.56       0    -6.49    -4.20
```

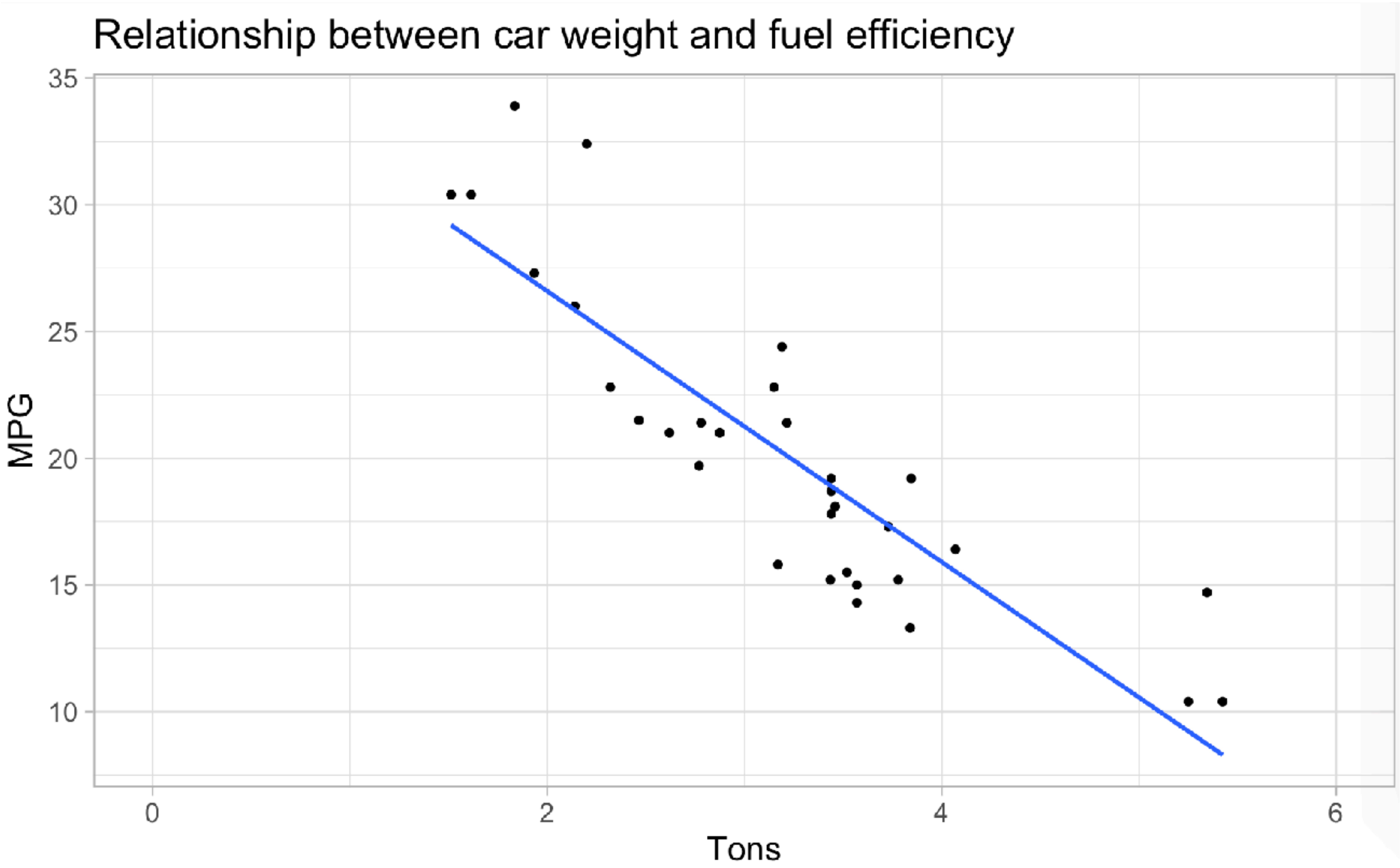


$$\hat{mpg} = 37.3 + -5.34 \times weight + \epsilon$$

```
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    37.3
2 wt           -5.34
```

NOT YET; end of semester

Why doesn't it look right?



Interpretation

A one unit increase in X is *associated* with a β_1 increase (or decrease) in Y, on average

$$\hat{mpg} = 37.3 + -5.34 \times weight + \epsilon$$

Put the above in human words