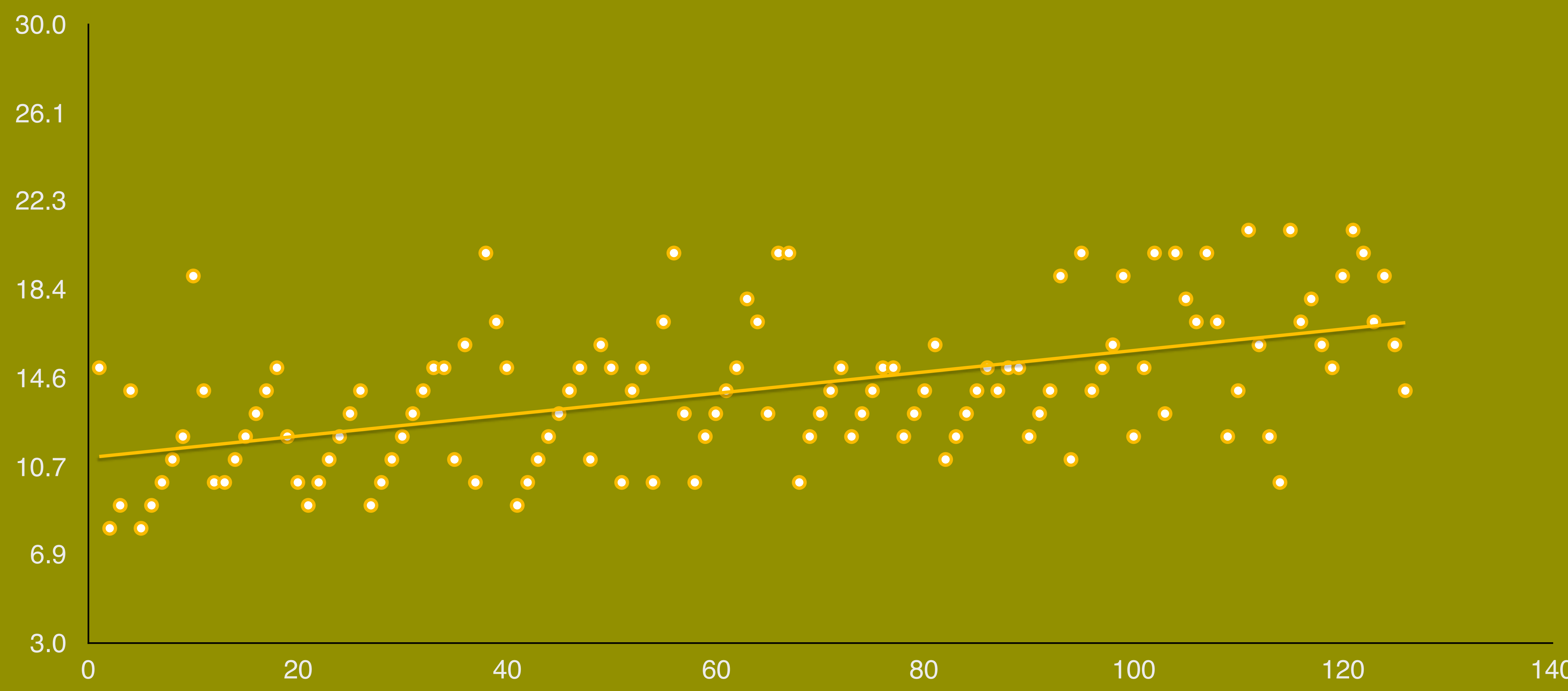
Linear Regression



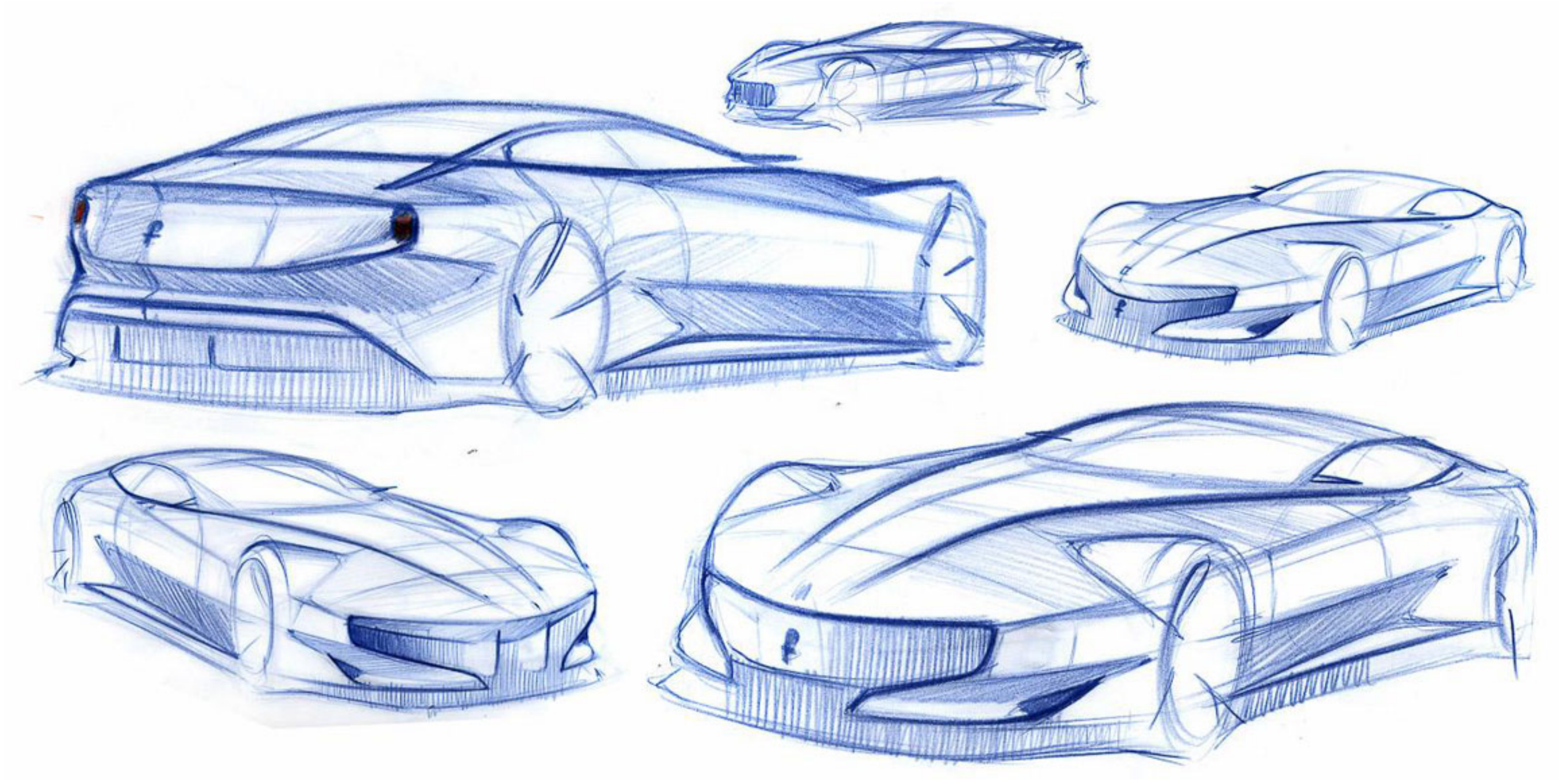
Instructors

Sourish Das, PhD
Associate Professor
Chennai Mathematical Institute

Regression

- What is the mileage of a prototype car?
- Why am I identifying this problem as possible regression problem?
- What is the target variable?
- Is it continuous variable?

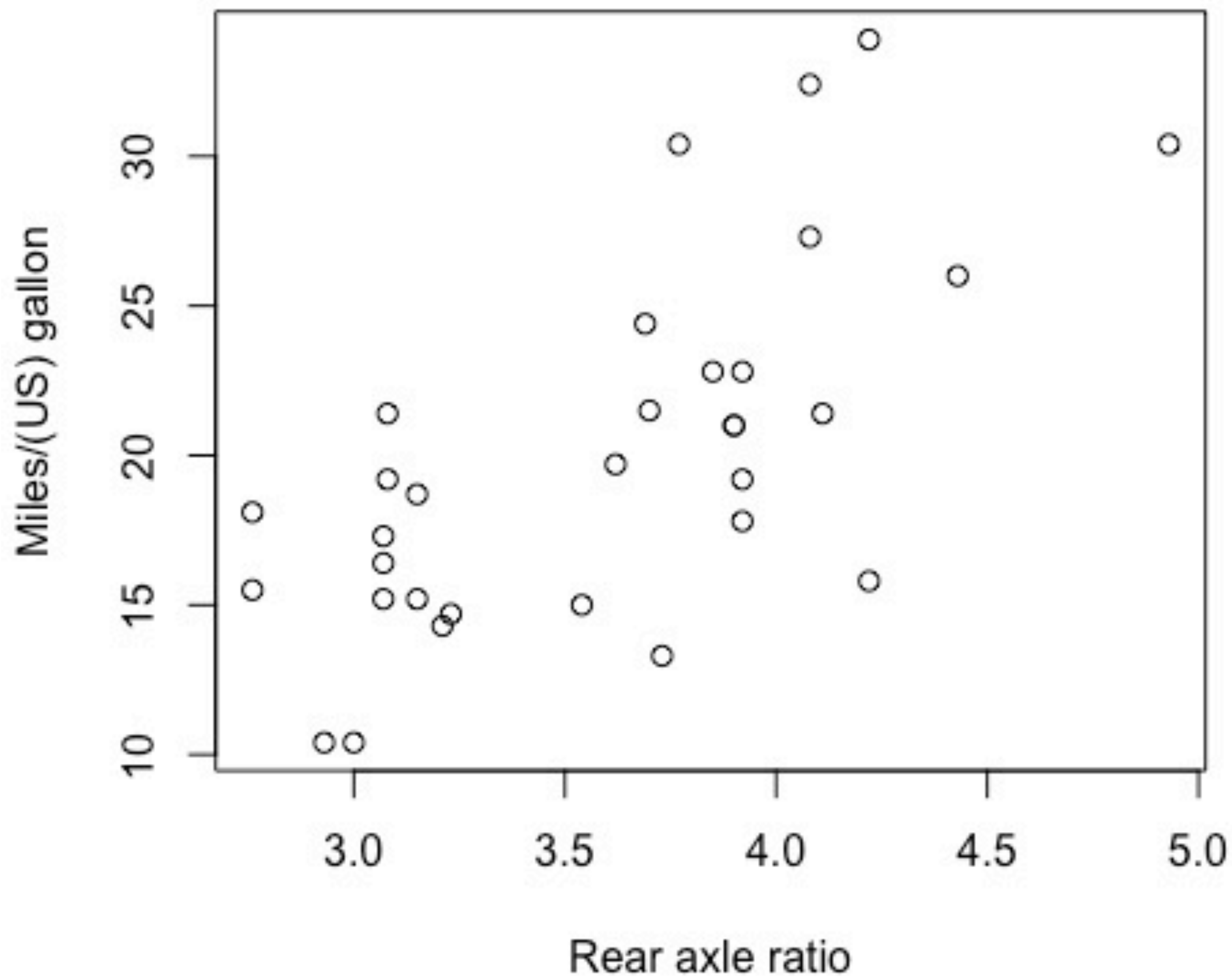
Regression



mtcars DataSet Available in R

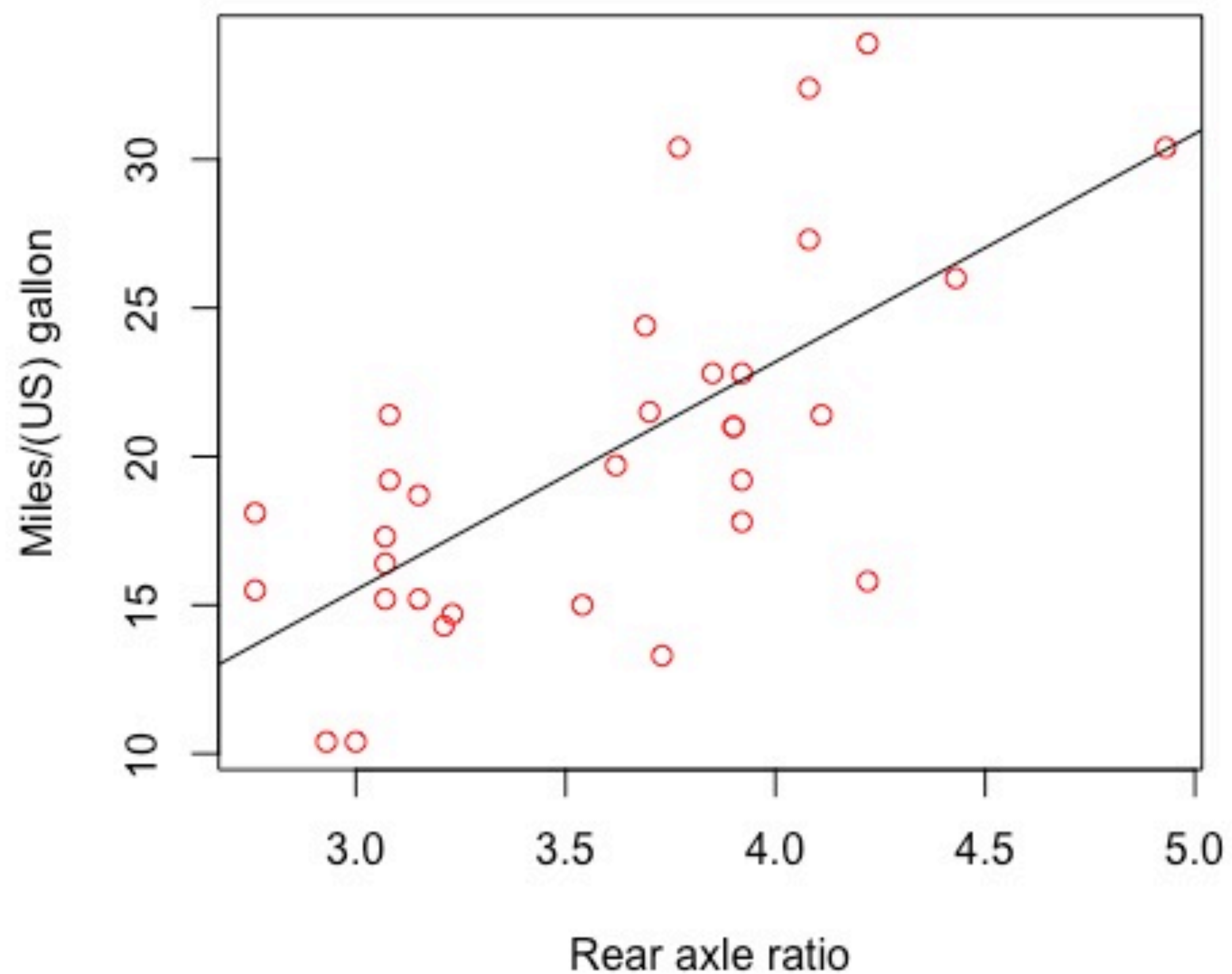
	mpg	cyl	drat	hp
Mazda RX4	21.0	6	3.90	110
Mazda RX4 Wag	21.0	6	3.90	110
Datsun 710	22.8	4	3.85	93
Hornet 4 Drive	21.4	6	3.08	110
....				
Prototype	?	4	3.90	120

Scatter Plot

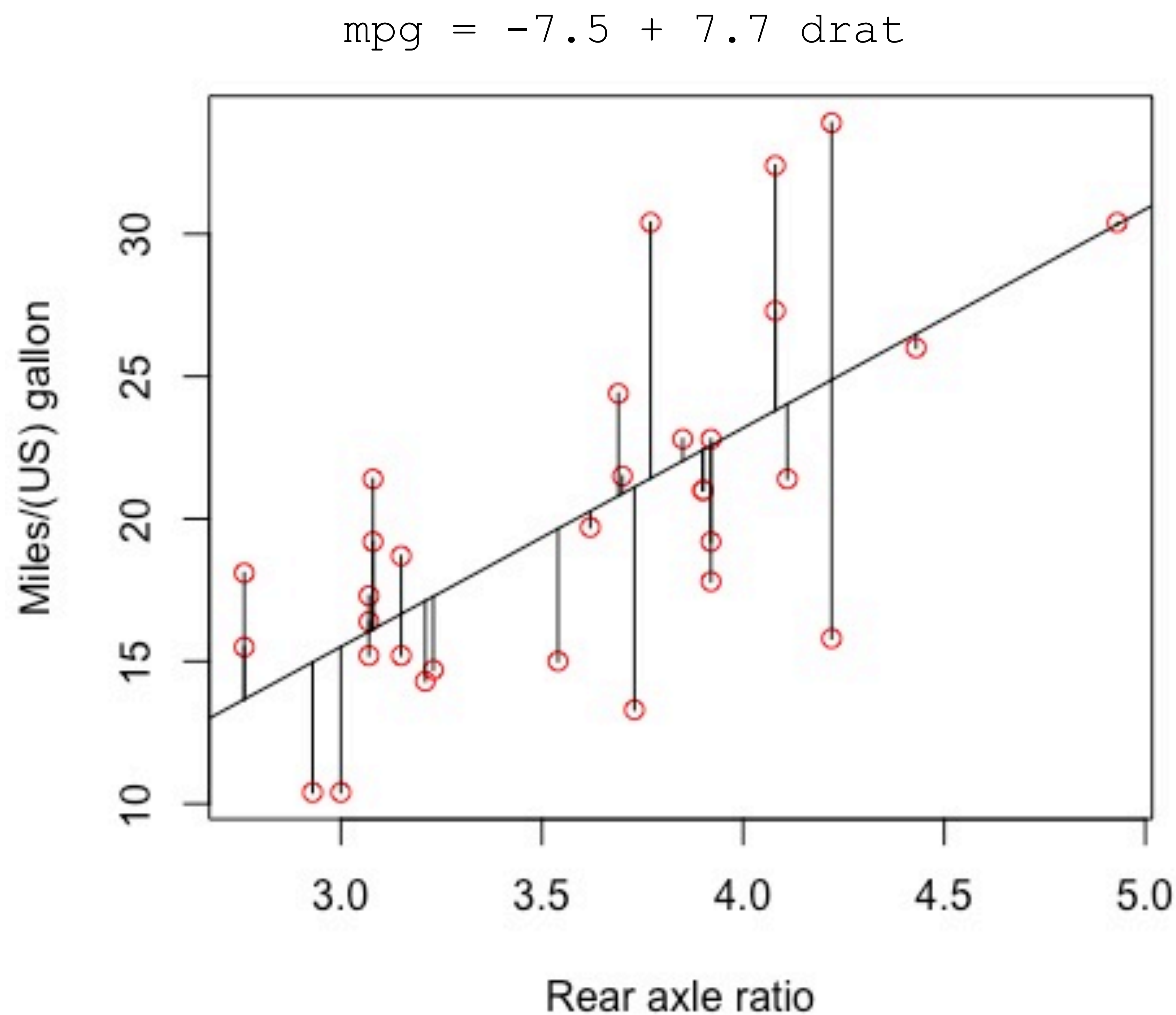


Fit A line

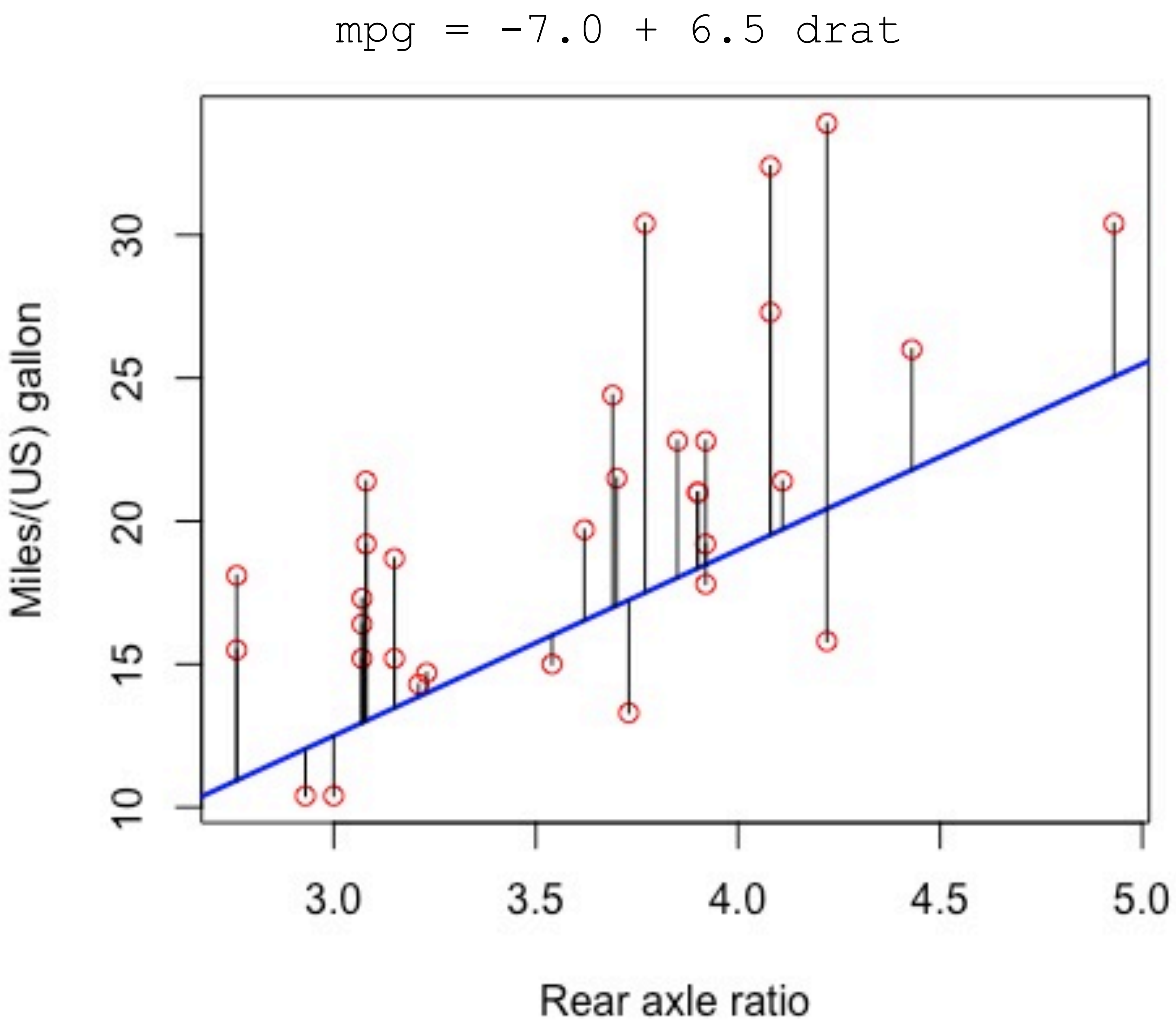
$$mpg = -7.5 + 7.7\ drat$$



Fit A line


mpg = -7.5 + 7.7 drat
Miles/(US) gallon
10
15
20
25
30
3.0
3.5
4.0
4.5
5.0
Rear axle ratio

Fit Another Line



Minimize The Error Sum Of Square

- Consider the model as:

$$y = a + b\,x + e$$

- Residual/Error sum of square :

$$RSS(a, b) = \sum_{i=1}^{n} (y - a - bx)^2$$

Minimize The Residual Sum of Square

- Differentiate **RSS(a, b)** with respect to **a** and **b**

$$\frac{\partial}{\partial a} RSS(a, b) = 0$$

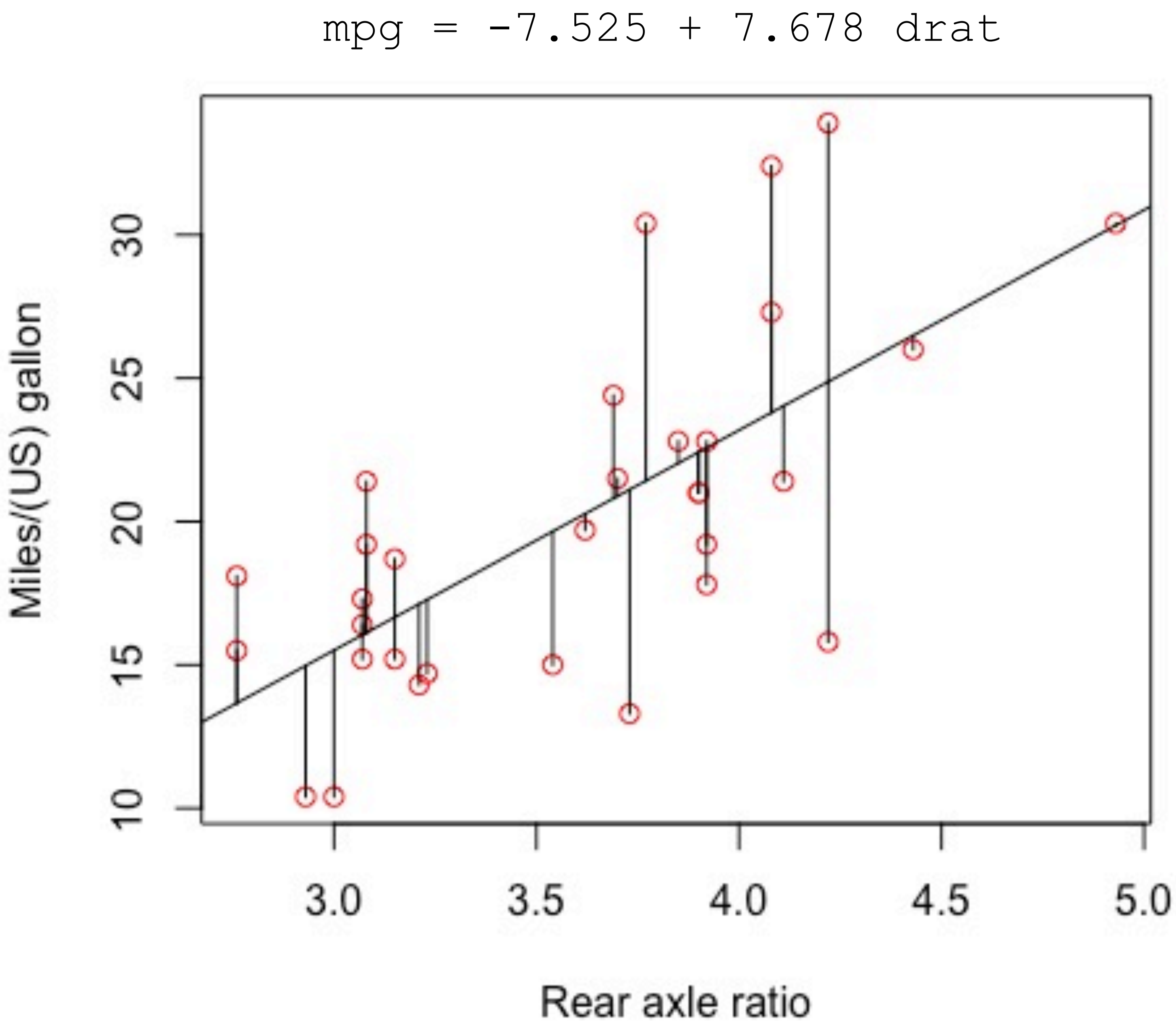
$$\frac{\partial}{\partial b} RSS(a, b) = 0$$

- Solution is known as Ordinary Least Square (OLS) estimates

Model fitting with R and Python

- One can fit linear regression to data very easily using R and/or Python
  - R has a built-in function called **"lm"** – you can use it
  - Python has **"lm.fit"** in the linear_model of SKLearn module

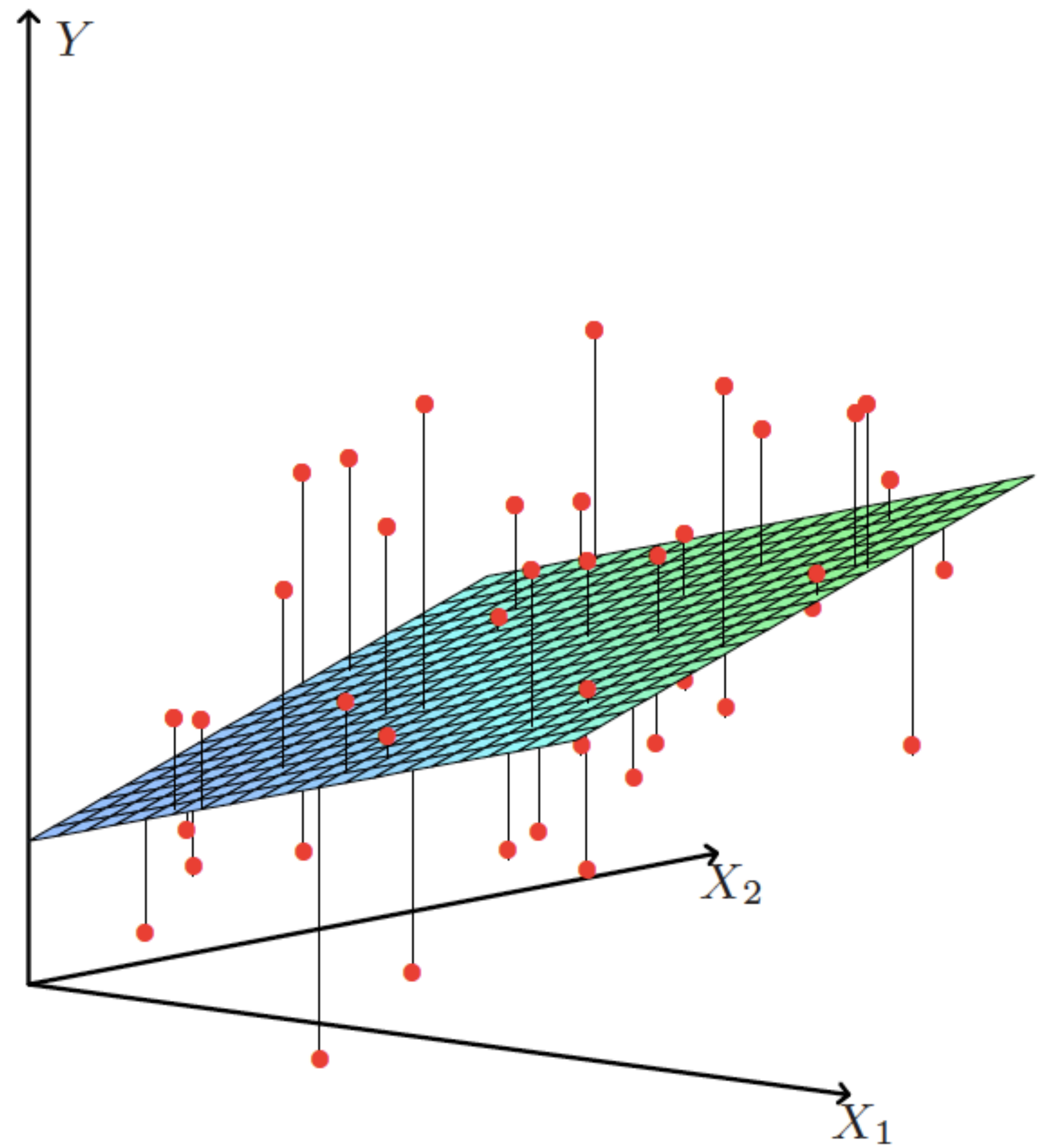
Best Fit With Minimum RSS



Regression With Multiple Features

- Regression model with two feature

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$



Estimating Coefficients

$$\text{mpg} = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

```
Coefficients:

               Estimate Std.Error t value   Pr(>|t|)
(Intercept)       10.79      5.08   2.125   0.042238 *
drat               4.70      1.19   3.943   0.000467 ***
hp                -0.05      0.01  -5.573   5.17e-06 ***

---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
Residual standard error: 3.17 on 29 degrees of freedom
Multiple R-squared:  0.7412,   Adjusted R-squared:
0.7233
F-statistic: 41.52 on 2 and 29 DF,  p-value: 3.081e-09
                                       Sample size = 32
```

Estimating Coefficients

$$\text{mpg} = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

Coefficients:

10.79 is the estimated value of β_0 from data

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	10.79	5.08	2.125	0.042238	*
drat	4.70	1.19	3.943	0.000467	***
hp	-0.05	0.01	-5.573	5.17e-06	***

---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.17 on 29 degrees of freedom
Multiple R-squared: 0.7412, Adjusted R-squared: 0.7233
F-statistic: 41.52 on 2 and 29 DF, p-value: 3.081e-09
Sample size = 32

Estimating Coefficients

$$\text{mpg} = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

Coefficients:

4.70 is the estimated value of β_1 from data

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	10.79	5.08	2.125	0.042238	*
drat	4.70	1.19	3.943	0.000467	***
hp	-0.05	0.01	-5.573	5.17e-06	***

---Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.17 on 29 degrees of freedom
Multiple R-squared: 0.7412, Adjusted R-squared: 0.7233
F-statistic: 41.52 on 2 and 29 DF, p-value: 3.081e-09

Sample size = 32

Estimating Coefficients

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

Coefficients:

-0.05 is the estimated value of β_2 from data

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	10.79	5.08	2.125	0.042238	*
drat	4.70	1.19	3.943	0.000467	***
hp	-0.05	0.01	-5.573	5.17e-06	***

---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.17 on 29 degrees of freedom

Multiple R-squared: 0.7412, Adjusted R-squared: 0.7233

F-statistic: 41.52 on 2 and 29 DF, p-value: 3.081e-09

Sample size = 32

Estimating Coefficients

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

```
Coefficients:

               Estimate Std.Error t value   Pr(>|t|)
(Intercept)       10.79      5.08   2.125   0.042238 *
drat               4.70      1.19   3.943   0.000467 ***
hp                -0.05      0.01  -5.573   5.17e-06 ***
```

So effectively our model is

Mpg = 10.79 + 4.70 drat -0.05 hp

Now if we know drat = 3.90 and hp = 120 for a prototype car then the expected mpg is:
Mpg = 10.79 + 4.70 * 3.90 -0.05 * 120 = 23.12 (approx)

Estimating Coefficients

$$\text{mpg} = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

```
Coefficients:

             Estimate Std.Error t value   Pr(>|t|)
(Intercept)     10.79      5.08   2.125   0.042238 *
drat             4.70      1.19   3.943   0.000467 ***
hp              -0.05      0.01  -5.573   5.17e-06 ***
```

We assume that data has some inherent randomness – which is beyond our control.

Because of this randomness the estimates of β_0, β_1, β_2 are prone to error. The standard error (Std.Error) provides an estimate of error associated with the estimates of β_0 , β_1, β_2.

Of course smaller the estimates -- better it is --

Finding t-value

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

```
Coefficients:

              Estimate Std.Error t value   Pr(>|t|)
(Intercept)      10.79      5.08   2.125   0.042238 *
drat              4.70      1.19   3.943   0.000467 ***
hp               -0.05      0.01  -5.573   5.17e-06 ***
```

The t-value is the ratio of Estimate/Std.Error

t-value for β_0 : 10.79/5.08 = 2.12
t-value for β_1 : 4.70/1.19 = 3.94
t-value for β_2 : 4.70/1.19 = -5.57

If the absolute value of t-value is large then it indicates the predictor has significant effect on the dependent variable.

Finding p-value

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

```
Coefficients:

               Estimate Std.Error t value   Pr(>|t|)
(Intercept)       10.79      5.08   2.125   0.042238 *
drat               4.70      1.19   3.943   0.000467 ***
hp                -0.05      0.01  -5.573   5.17e-06 ***

---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Pr(>|t|) is the known as the p-value. The p-value is a probability value. The p-value will be always between 0 and 1.
Lower p-value indicates statistically significant effect of predictor on dependent variable.
The '***' indicates the p-value is in between 0 and 0.001

Multiple R-square

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

Coefficients:

What is Multiple R-square ?

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	10.79	5.08	2.125	0.042238	*
drat	4.70	1.19	3.943	0.000467	***
hp	-0.05	0.01	-5.573	5.17e-06	***

---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.17 on 29 degrees of freedom

Multiple R-squared: 0.7412, Adjusted R-squared: 0.7233

F-statistic: 41.52 on 2 and 29 DF, p-value: 3.081e-09

Sample size = 32

Multiple R-square

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

```
Coefficients:

              Estimate Std.Error t value   Pr(>|t|)
(Intercept)      10.79      5.08   2.125   0.042238 *
drat              4.70      1.19   3.943   0.000467 ***
hp               -0.05      0.01  -5.573   5.17e-06 ***

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

In popular term, “Multiple R-Square” indicates what percentage of variation of the target variable is being explained by the model

$$R^2 = cor(y, \hat{y})^2, \text{ where } \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Residual Standard Error

$$mpg = \beta_0 + \beta_1 drat + \beta_2 hp + e$$

Coefficients:

What is Residual Standard error ?

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	10.79	5.08	2.125	0.042238	*
drat	4.70	1.19	3.943	0.000467	***
hp	-0.05	0.01	-5.573	5.17e-06	***

---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.17 on 29 degrees of freedom

Multiple R-squared: 0.7412, Adjusted R-squared: 0.7233

F-statistic: 41.52 on 2 and 29 DF, p-value: 3.081e-09

Sample size = 32

What Is Residual Standard Error?

- The Residual standard error helps us to estimate the confidence interval for the predicted value from the model
- Suppose we know **drat = 3.90** and **hp = 120** for a prototype car then the expected mpg is:
- mpg = 10.79 + 4.70 * 3.90 -0.05 * 120 = 23.12 (approx)
- Not necessarily our final value of mpg will be exactly 23.12 – it will be some what plus or minus.
- So we can give an upper-bound and lower-bound for our mpg prediction

What Is Residual Standard Error?

- The residual standard error is 3.17
- Our prediction for mpg is 23.12
- So lower bound for our prediction is 23.12 – 2*3.17 = 16.78
- And upper bound for our prediction is 23.12 + 2*3.17 = 29.46
- So we can say that with 95% confidence that the final realized value of the mpg of the prototype car will be between 16.78 and 29.46 i.e. (16.78 , 29.46)

Please watch the session video on Linear Regression to have better clarity on the topic.

Thank You