Simple & Multiple Linear Regression Modelling and Interpretation

Vinay Kulkarni

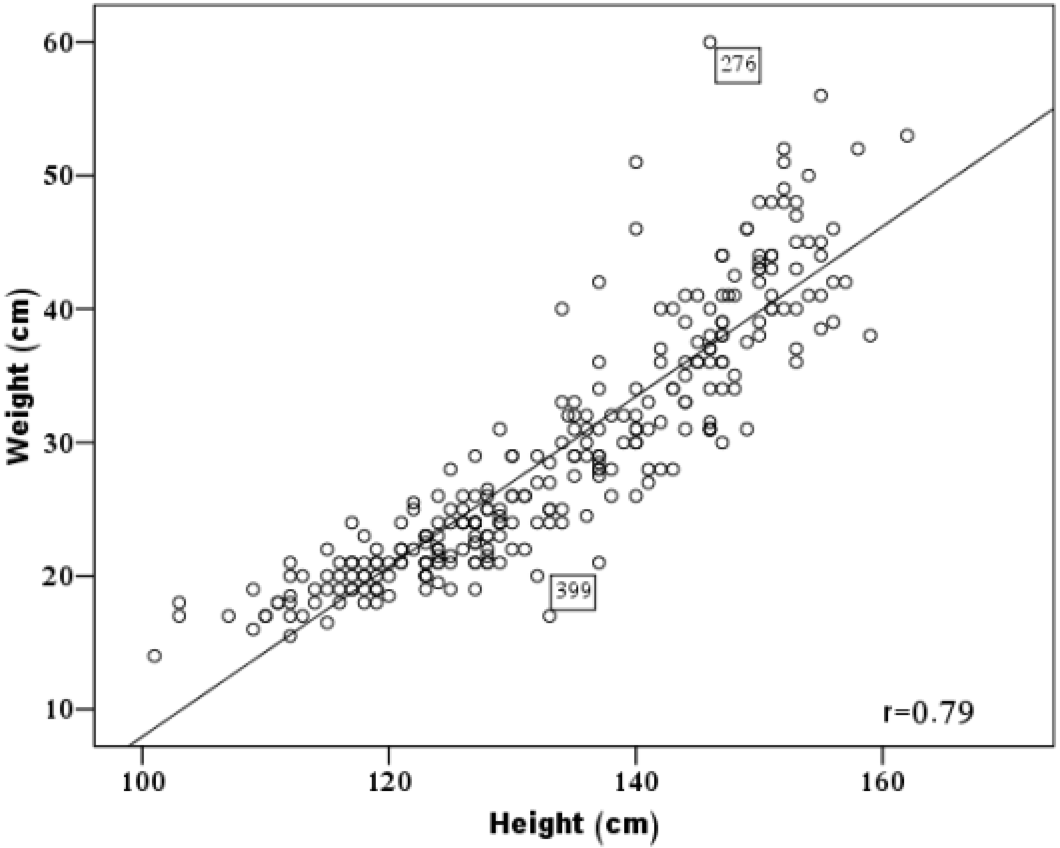
Regression Analysis

- Are two given variables related to each other?
- Can we find a linear relationship between them?
- Can we predict the value of a random variable?
- These questions can be answered through Regression Analysis
- Regression
  - "Going back to mediocrity or average"

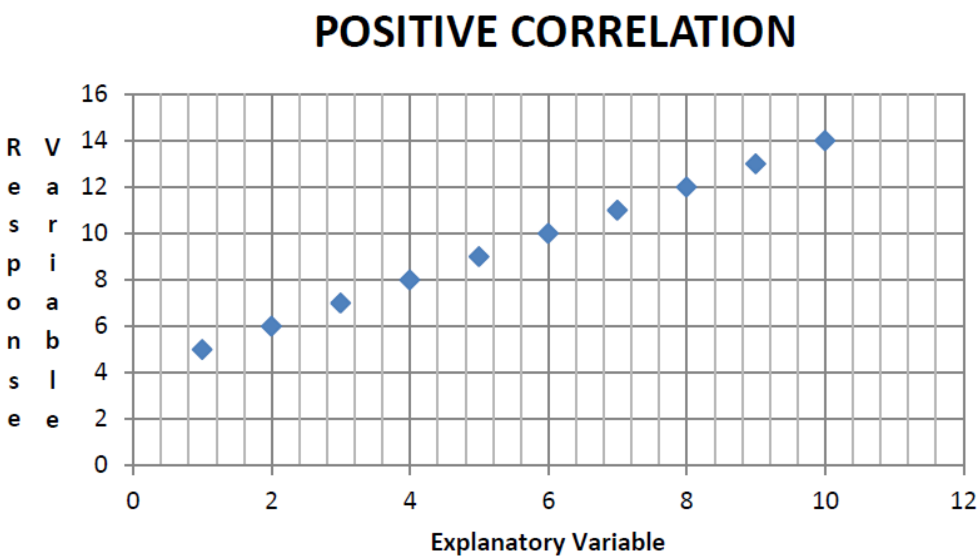
Regression Analysis

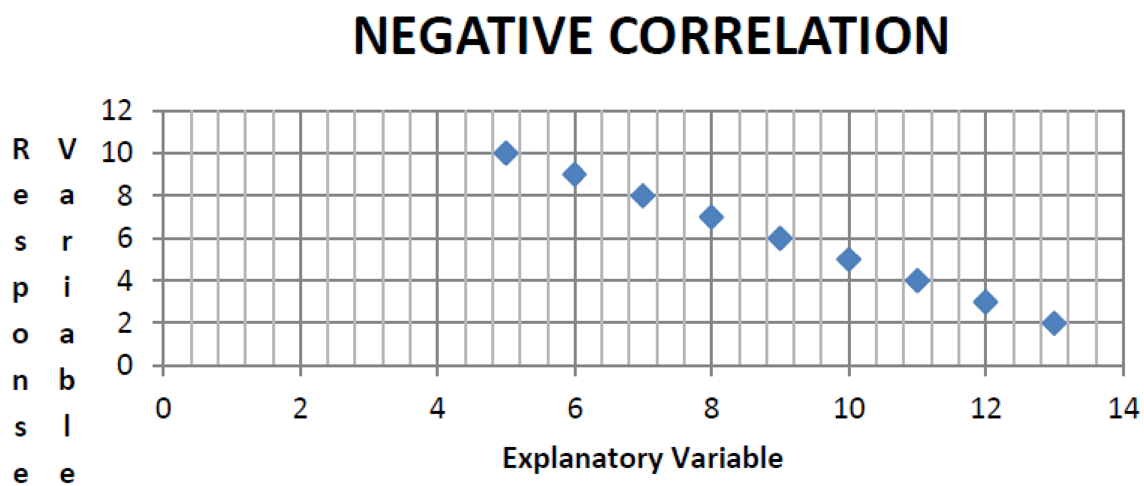
- Explanatory or Predictor or Independent variable(s)
- Response or Dependent variable
  - It is not always clear which variable is **predictor**, and which is the **response**!
- Types of regression analysis
  - Linear Regression
    - One Predictor, one response variable
  - Multiple Regression
    - Many Predictors, one response variable
- Other types of regressions:
  - Logarithmic, Exponential, Quadratic, Cubic

First use Scatter Plot to assess the relationship



First step: Use scatter plot





Simple Regression Model

- Simple regression model is of the form
  - $Y = \beta_0 + \beta_1 x + e$
- Where
  - Y = Response variable
  - X = Regressor or explanatory variable
  - β_0 is the y intercept
  - β_1 is regression coefficient or slope
  - e is the random error
    - Has normal distribution
      - With mean 0
      - With variance σ^2

Least Squares Regression Line

Least squares regression line $\hat{y} = b_0 + b_1 x$

It passes through the '**mean**' point

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Each point of the sample satisfies

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Where

$$e_i = y_i - \hat{y}_1$$

Is the residual error, the square of which is minimized

Least Squares Regression Line

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \mathrm{b}_0 - \mathrm{b}_1 \mathrm{x}_i)^2$$

The values of coefficients are arrived at by minimizing the SSE. The results are as follows:

$$b_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \right)}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \right)}, \text{ and} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

Expressions derived in a separate document

Partitioning the total variation

- After doing a (simple linear) regression analysis the variations in each value of the response variable (Y) can be tabulated as follows
  - Total variation = Variation explained by regression + Random variation
- Variation explained by regression
  - Attributable cause
- Random variation
  - Non-attributable causes
- If variation <u>explained by regression</u> is much higher than random variation
  - The response variable is said to be correlated to the explanatory variable

Partitioning the total variation

- Total error = distance of point from the mean =
  - Distance of actual point from point on line +
  - Distance of point on line from mean
- If the distance of actual point from point on line is small
  - ➔ Fitness is good
- The measure of goodness of fit is known as
  - Coefficient of determination: R^2
  - Closer this value is to 1, better the fit

Partitioning of the Total Variation

$Y_i = \bar{y} + (\hat{y}_i - \bar{y}) + (Y_i - \hat{y}_i)$, for i =1, 2, 3, …, n.

$Y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (Y_i - \hat{y}_i)$, for i = 1, 2, 3, …, n.

Total deviation = Deviation due to regression + Deviation about regression

$$\sum_{i=1}^{n}(Y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$$

$$SS_{\text{total}} \quad = \quad SS_{\text{regr}} + SS_{\text{residuals}}$$

$$SS_{\text{total}} = \sum_{i=1}^{n}(Y_i - \bar{y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\sum_{i=1}^{n} Y_i^2}{n}$$

$$SS_{\text{regr}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_0 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} Y_i^2}{n}, \text{ and}$$

$$SS_{\text{residuals}} = SS_{\text{total}} - SS_{\text{regr}} = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i .$$

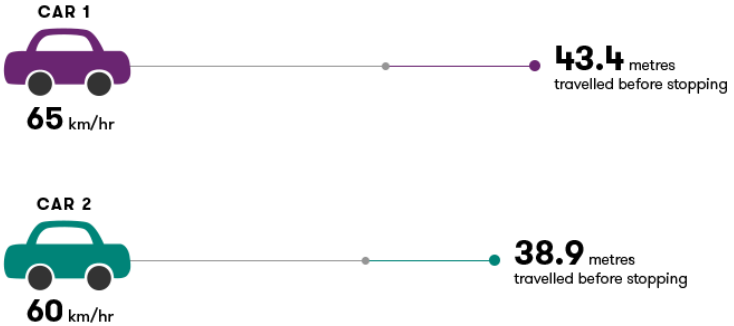
$r^2 = SS_{\text{regr}} / SS_{\text{total}}$

Simple Linear Regression: Cars - Breaking Distance

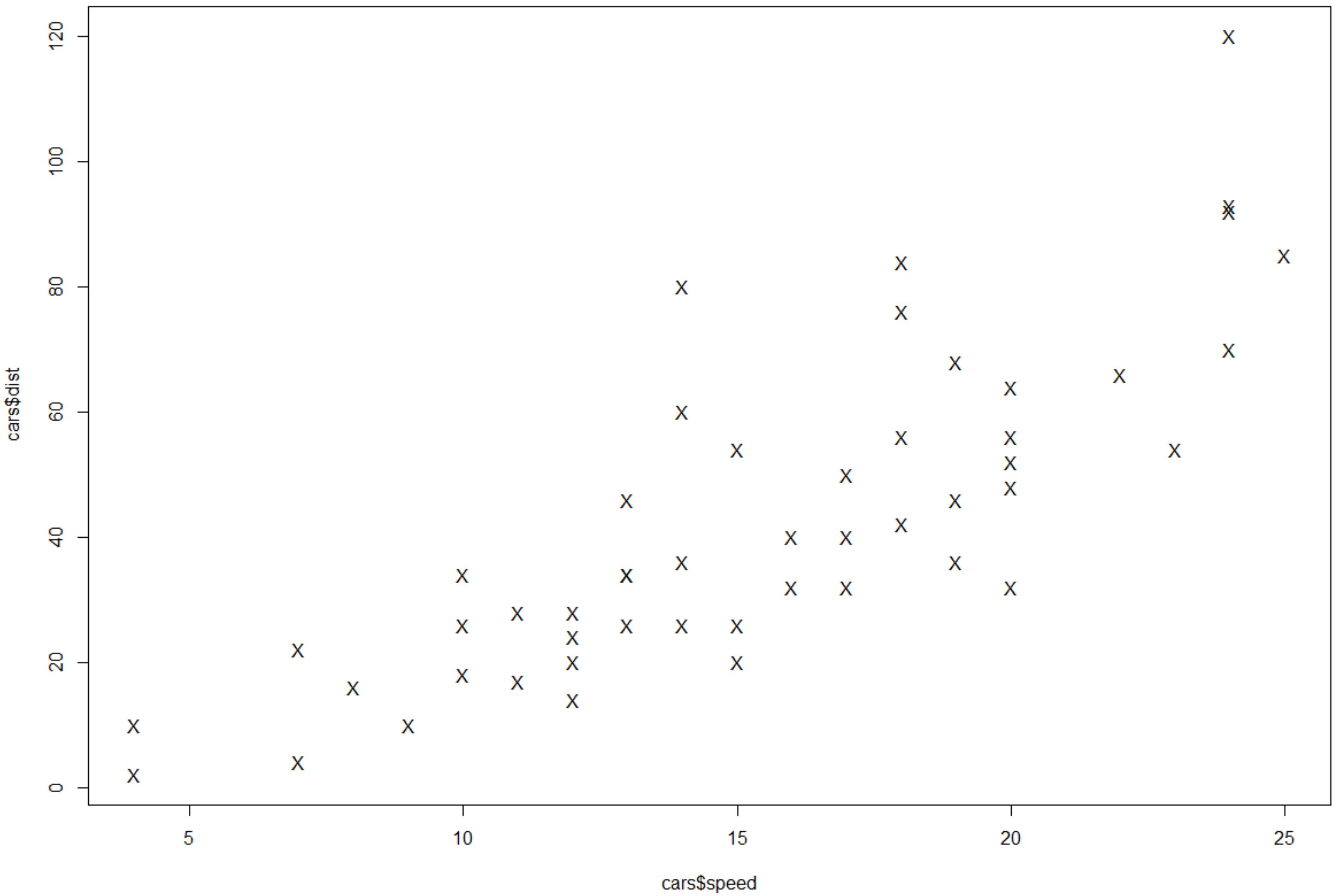
- Data : Speed v/s Breaking Distance

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28
12	12	14
13	12	20
14	12	24
15	12	28
16	13	26
17	13	34
18	13	34
19	13	46
20	14	26
21	14	36
22	14	60
23	14	80
24	15	20
25	15	26

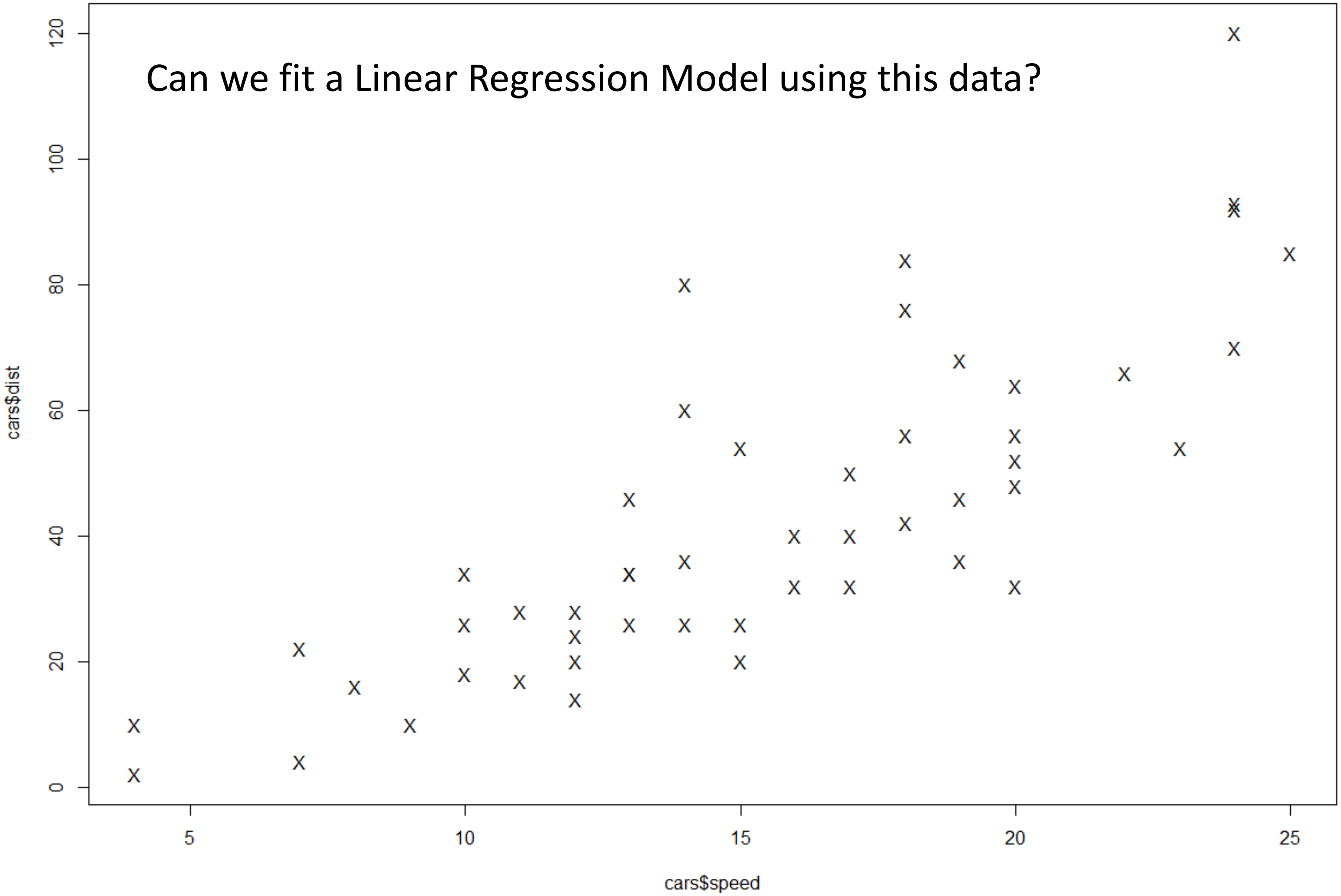
	speed	dist
26	15	54
27	16	32
28	16	40
29	17	32
30	17	40
31	17	50
32	18	42
33	18	56
34	18	76
35	18	84
36	19	36
37	19	46
38	19	68
39	20	32
40	20	48
41	20	52
42	20	56
43	20	64
44	22	66
45	23	54
46	24	70
47	24	92
48	24	93
49	24	120
50	25	85



First step: Visualization



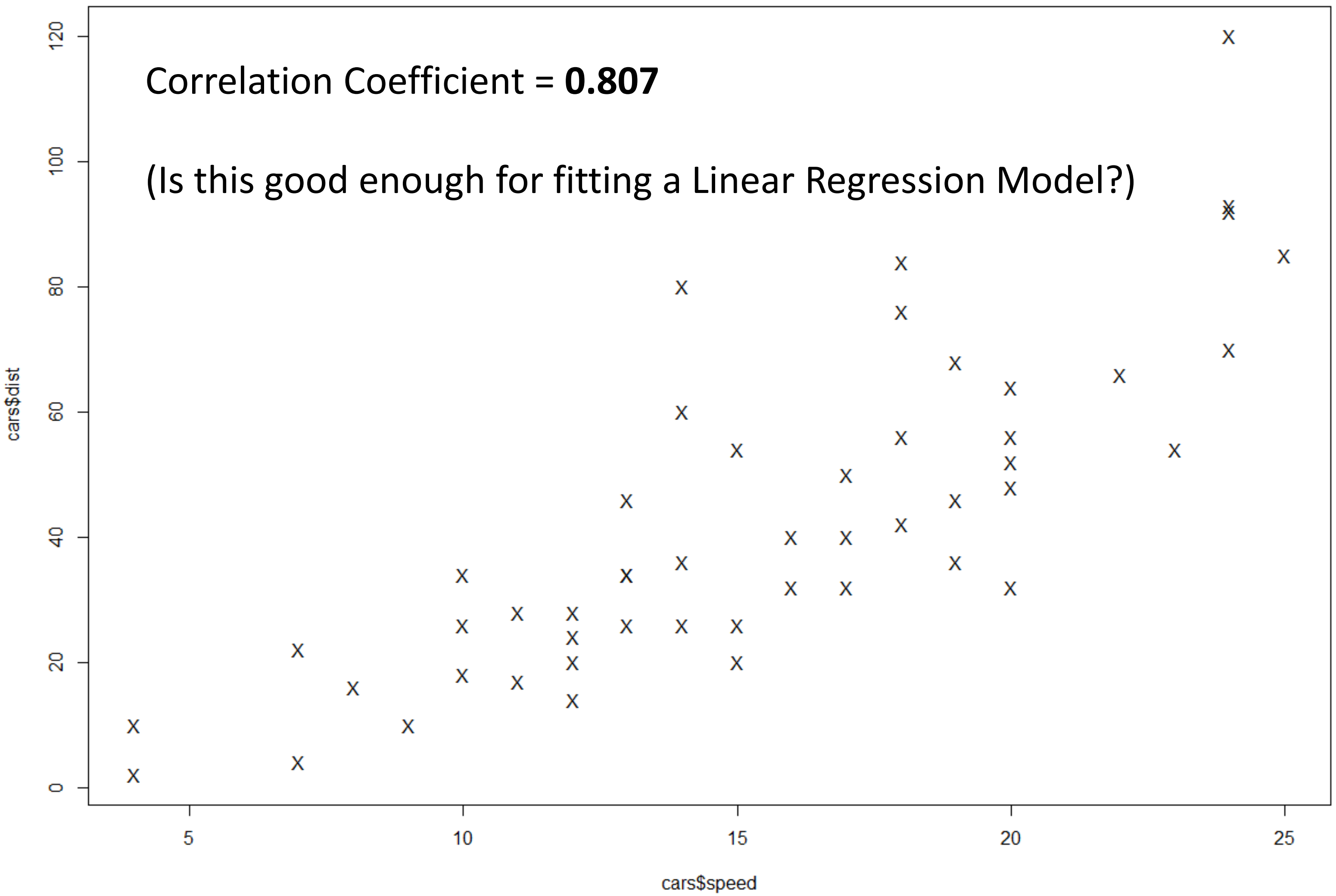
Plot: Speed v/s Distance



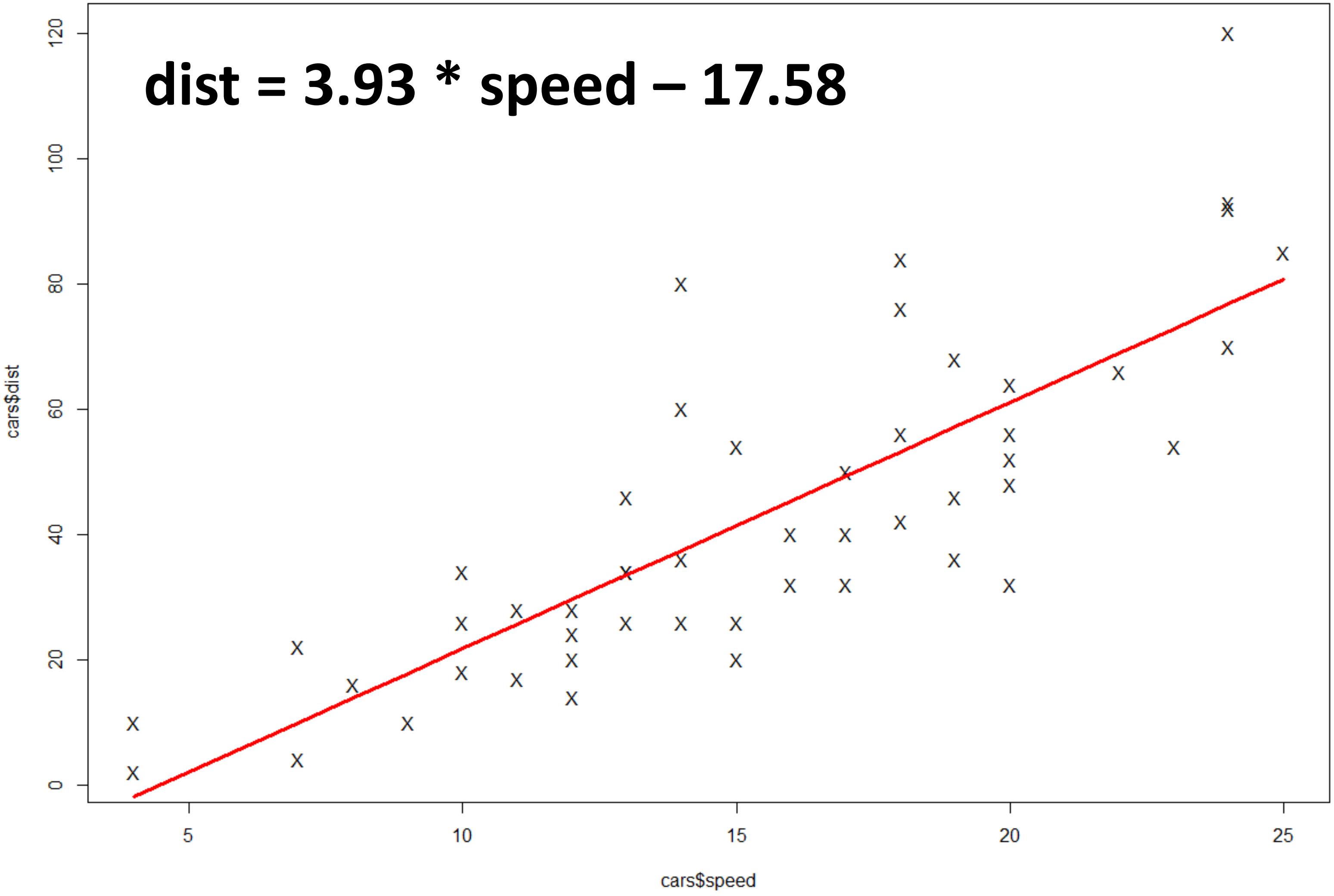
Correlation Coefficient

- Is there a good enough reason to fit a Linear Regression model?
  - We need to check if the two variables in general have a linear relationship
  - Visual check is one method: works when only one predictor variable is involved
  - Finding out the correlation coefficient is another quick check
  - Correlation Coefficient: $r_{xy} = \frac{1}{n} \sum \frac{(x_i - \bar{x})}{\sigma_x} \cdot \frac{(y_i - \bar{y})}{\sigma_y}$
  - Also: $\beta_1 = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$

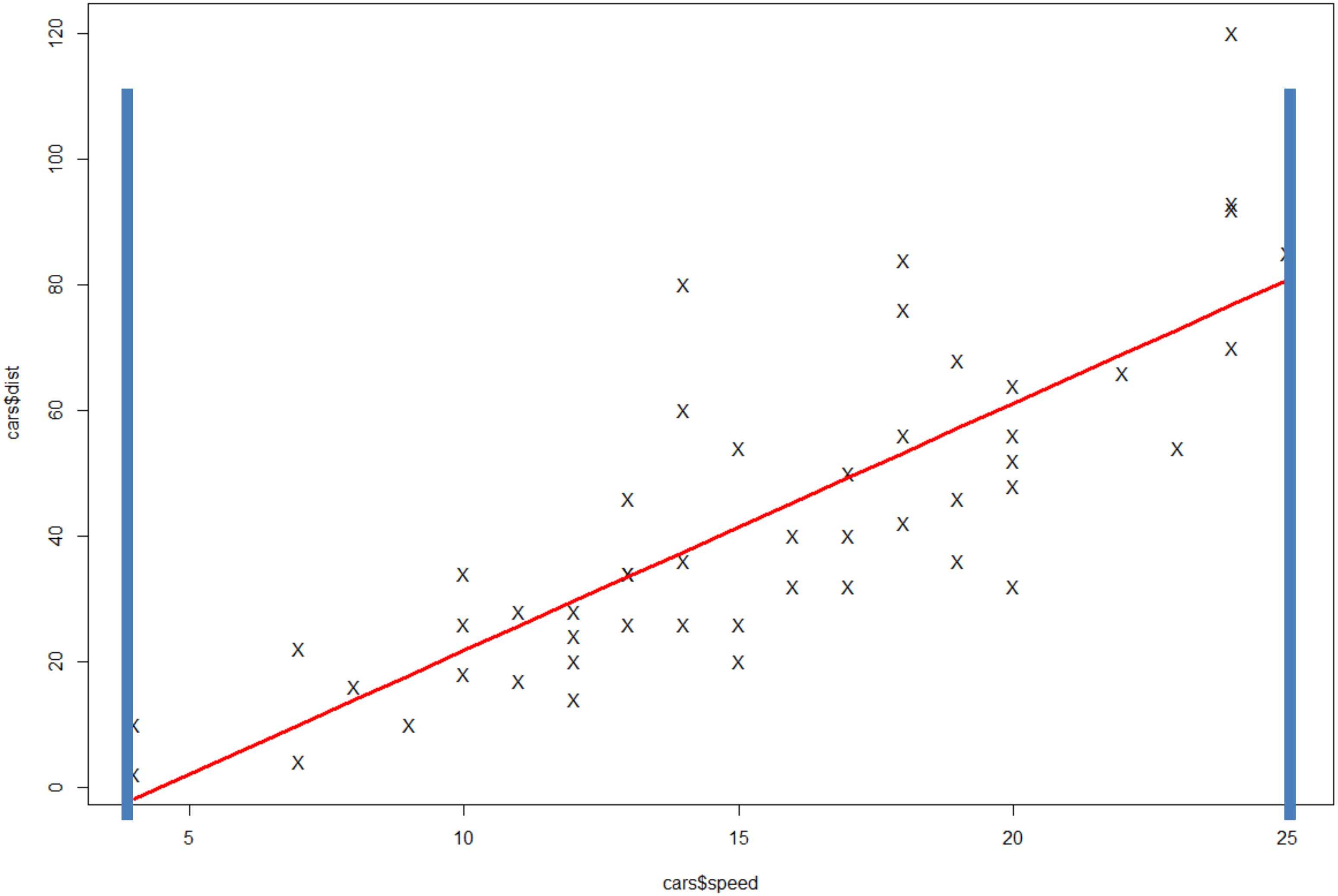
Plot: Speed v/s Distance



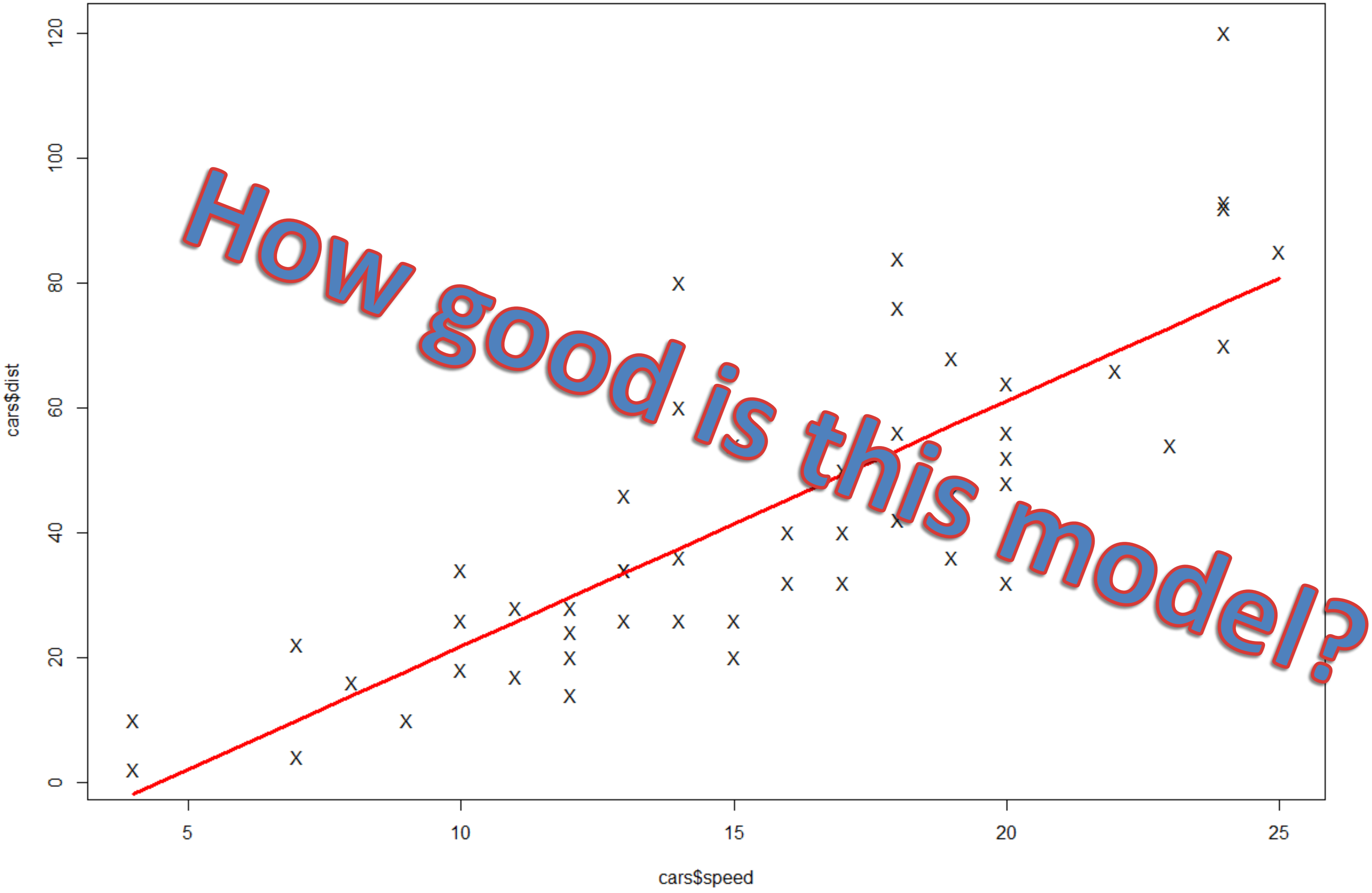
Linear Regression Model (based on LSE)



Linear Regression Model : Validity



Linear Regression Model:



How good is the model?

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Judged by analyzing various information generated during the process of Linear Regression

Understanding the model: Residuals

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

RESIDUALS

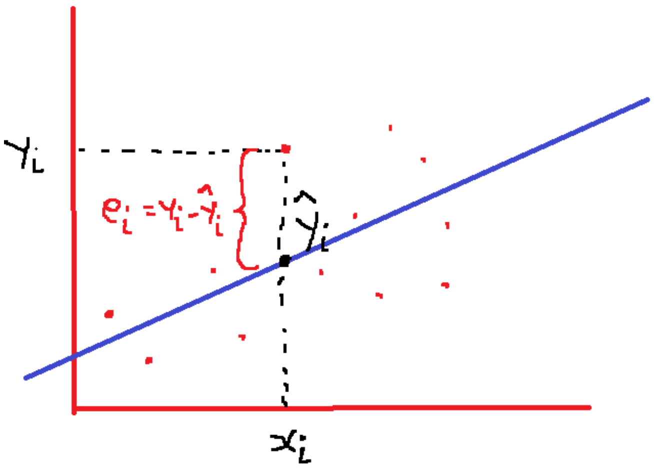
Understanding the model: Residuals

- Residuals or Errors

$$e_i = (y_i - \hat{y}_i)$$

- Residuals (Cars example)

1	2	3	4	5	6
3.849460	11.849460	-5.947766	12.052234	2.119825	-7.812584
7	8	9	10	11	12
-3.744993	4.255007	12.255007	-8.677401	2.322599	-15.609810
13	14	15	16	17	18
-9.609810	-5.609810	-1.609810	-7.542219	0.457781	0.457781
19	20	21	22	23	24
12.457781	-11.474628	-1.474628	22.525372	42.525372	-21.407036
25	26	27	28	29	30
-15.407036	12.592964	-13.339445	-5.339445	-17.271854	-9.271854
31	32	33	34	35	36
0.728146	-11.204263	2.795737	22.795737	30.795737	-21.136672
37	38	39	40	41	42
-11.136672	10.863328	-29.069080	-13.069080	-9.069080	-5.069080
43	44	45	46	47	48
2.930920	-2.933898	-18.866307	-6.798715	15.201285	16.201285
49	50				
43.201285	4.268876				



- Sum of errors = 0 (why?)
- Sum of square of errors = SSE = $\sum e_i^2 = 11353.52$

Understanding the model: Residuals

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

RESIDUALS

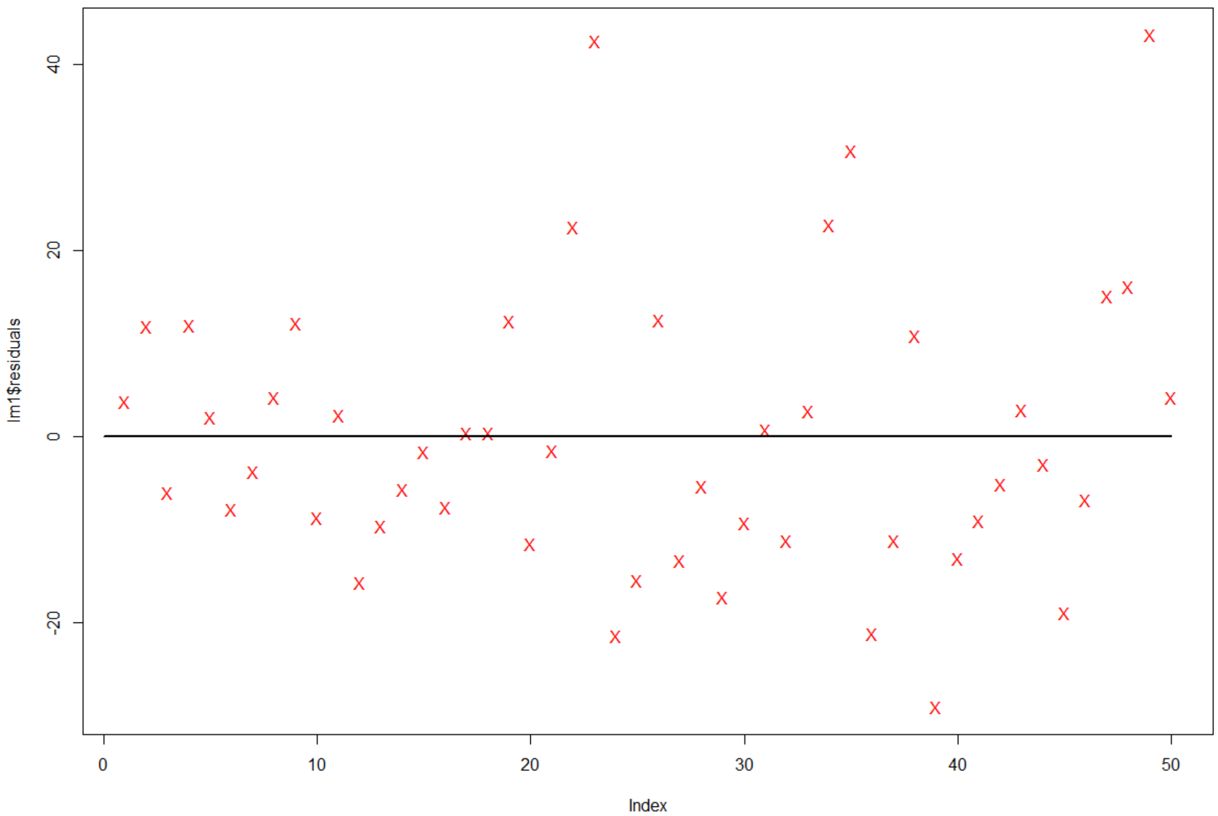
RESIDUAL STANDARD ERROR

$$S = \sqrt{\frac{1}{(n-2)} \cdot \sum e_i^2} = \sqrt{\frac{1}{(50-2)} \cdot 11353.52}$$

$$= 15.38$$

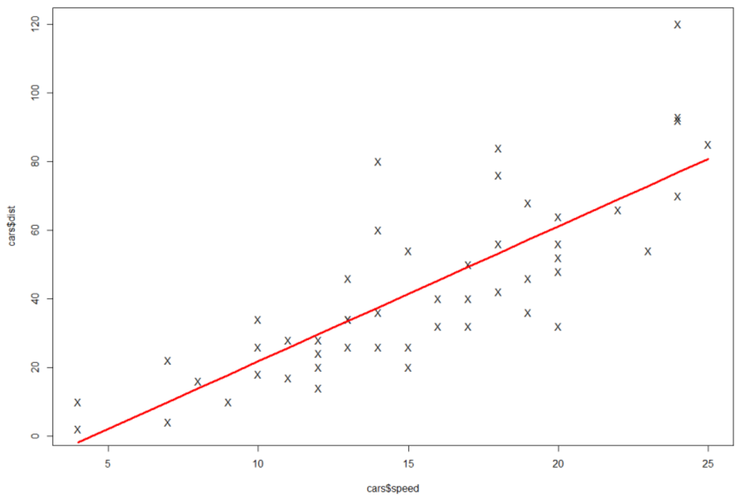
Interpreting Residuals …

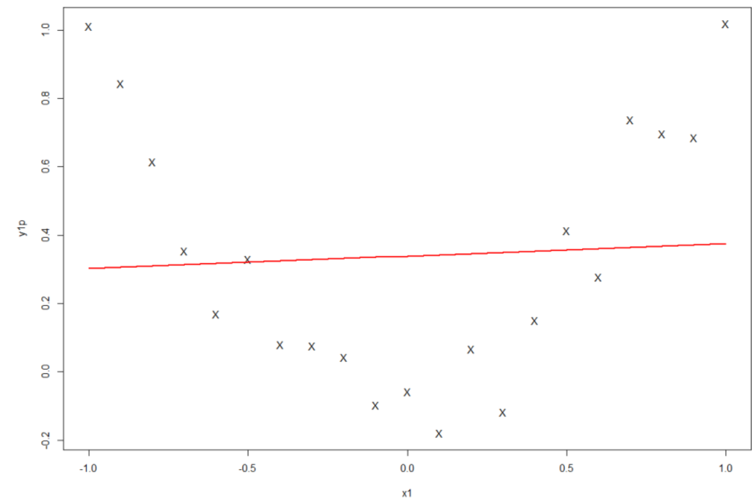
- Are they systematic or are they random?
- Good regression 🡐 Random Residuals

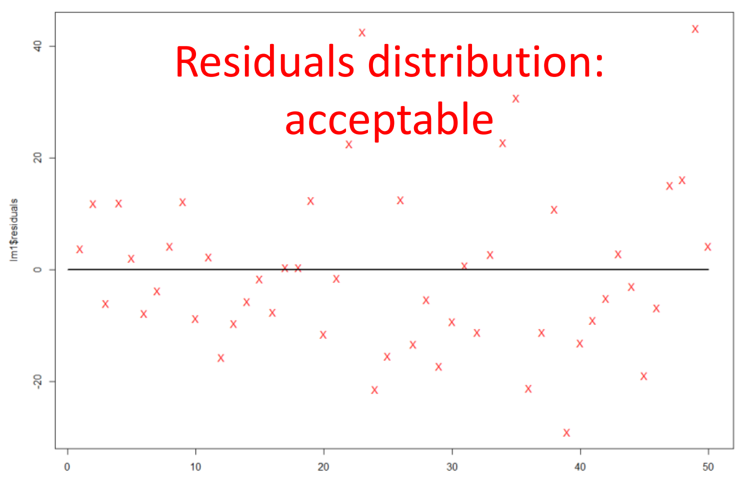


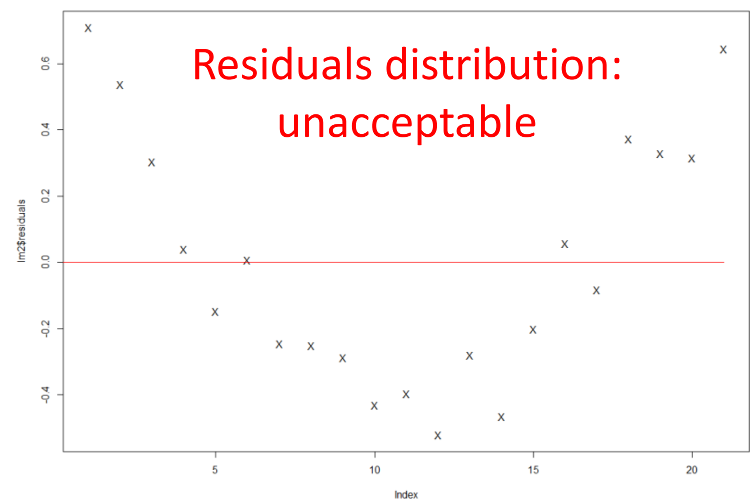
Interpreting Residuals …

- Are they systematic or are they random?
- Good regression 🡐 Random Residuals









Estimates of the Coefficients

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$\beta_0 = -17.5791$

$\beta_1 = 3.9324$

$$\beta_0 = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,.\,\overline{xy}}{(\overline{x^2} - \bar{x}^2)}$$

$$\beta_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{(\overline{x^2} - \bar{x}^2)} = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$$

where r_{xy} = Correlation Coefficient

Coefficient Standard Errors

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

S = RESIDUAL STD. ERROR

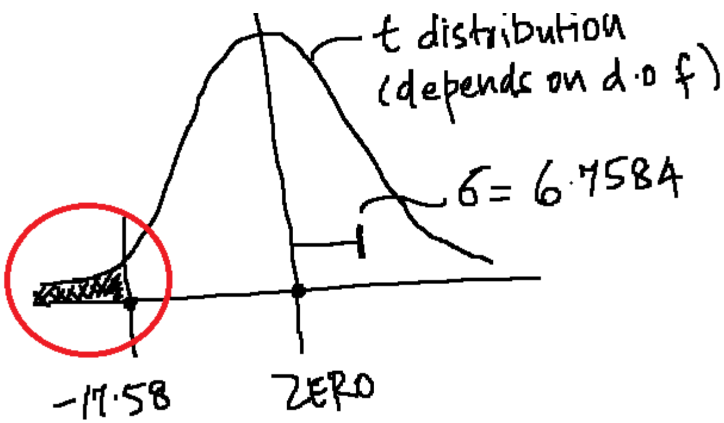
$$SE_{\beta_0} = \frac{S}{\sqrt{n}} \cdot \sqrt{1 + \frac{\bar{x}^2}{\sigma_x^2}} \qquad SE_{\beta_1} = \frac{S}{\sqrt{n}} \cdot \frac{1}{\sigma_x}$$

t-value and p-value

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```



$$t_{\beta_0} = \frac{-17.58}{6.7584} = -2.601$$

This should be less than 0.025

AREA UNDER THE PDF CURVE = 0.00615

$\therefore 2 \times 0.00615 = \underbrace{0.0123}$

This should be less than 0.05

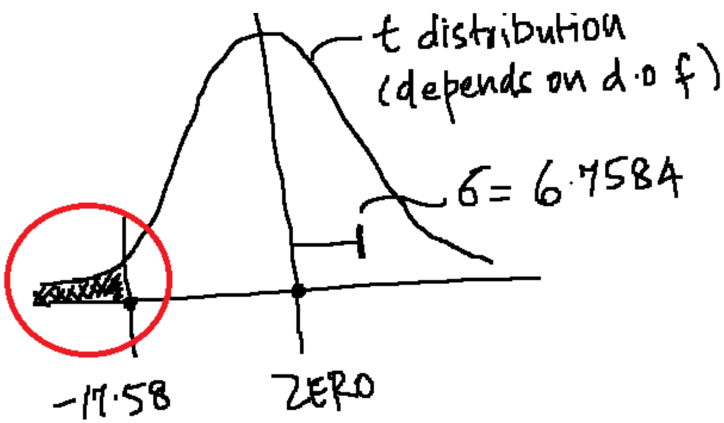
t-value and p-value

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Derived value of a coefficient is acceptable iff it's corresponding p-val is less than 0.05



$$t_{\beta_0} = \frac{-17.58}{6.7584} = -2.601$$

This should be less than 0.025

AREA UNDER THE PDF CURVE = 0.00615

$\therefore 2 \times 0.00615 = \underbrace{0.0123}_{}$

This should be less than 0.05

R-squared

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$Y_i = \bar{y} + (\hat{y}_i - \bar{y}) + (Y_i - \hat{y}_i)$, for i =1, 2, 3, …, n.

$Y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (Y_i - \hat{y}_i)$, for i = 1, 2, 3, …, n.

Total deviation = Deviation due to regression + Deviation about regression

$$\sum_{i=1}^{n} (Y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2$$

$SS_{total} \quad = \quad SS_{regr} + SS_{residuals}$

$r^2 = \quad SS_{regr} / SS_{total}$.

R-squared

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

R-squared should be close to 1 for the regression to be considered good

$$Y_i = \bar{y} + (\hat{y}_i - \bar{y}) + (Y_i - \hat{y}_i), \text{ for i =1, 2, 3, …, n.}$$

$$Y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (Y_i - \hat{y}_i), \text{ for i = 1, 2, 3, …, n.}$$

Total deviation = Deviation due to regression + Deviation about regression

$$\sum_{i=1}^{n} (Y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2$$

$SS_{total} = SS_{regr} + SS_{residuals}$

$r^2 = SS_{regr} / SS_{total}$.

$SS_{regr} = 21185.46$

$SS_{total} = 32538.98$

$r^2 = 0.6510$

$r = 0.8068$

= correlation coeff

Multiple Linear Regression

- **Multiple Linear Regression**. **Multiple linear regression** attempts to model the relationship between two or more explanatory variables and a response variable by fitting a **linear** equation to observed data (http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

Coefficients of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

Can be written as : $\mathbf{y} = \mathbf{X}\beta + \epsilon$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots x_{1k} \\ 1 & x_{21} & x_{22} \cdots x_{2k} \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 1 & x_{n1} & x_{n2} \cdots x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Coefficients of Multiple Linear Regression

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

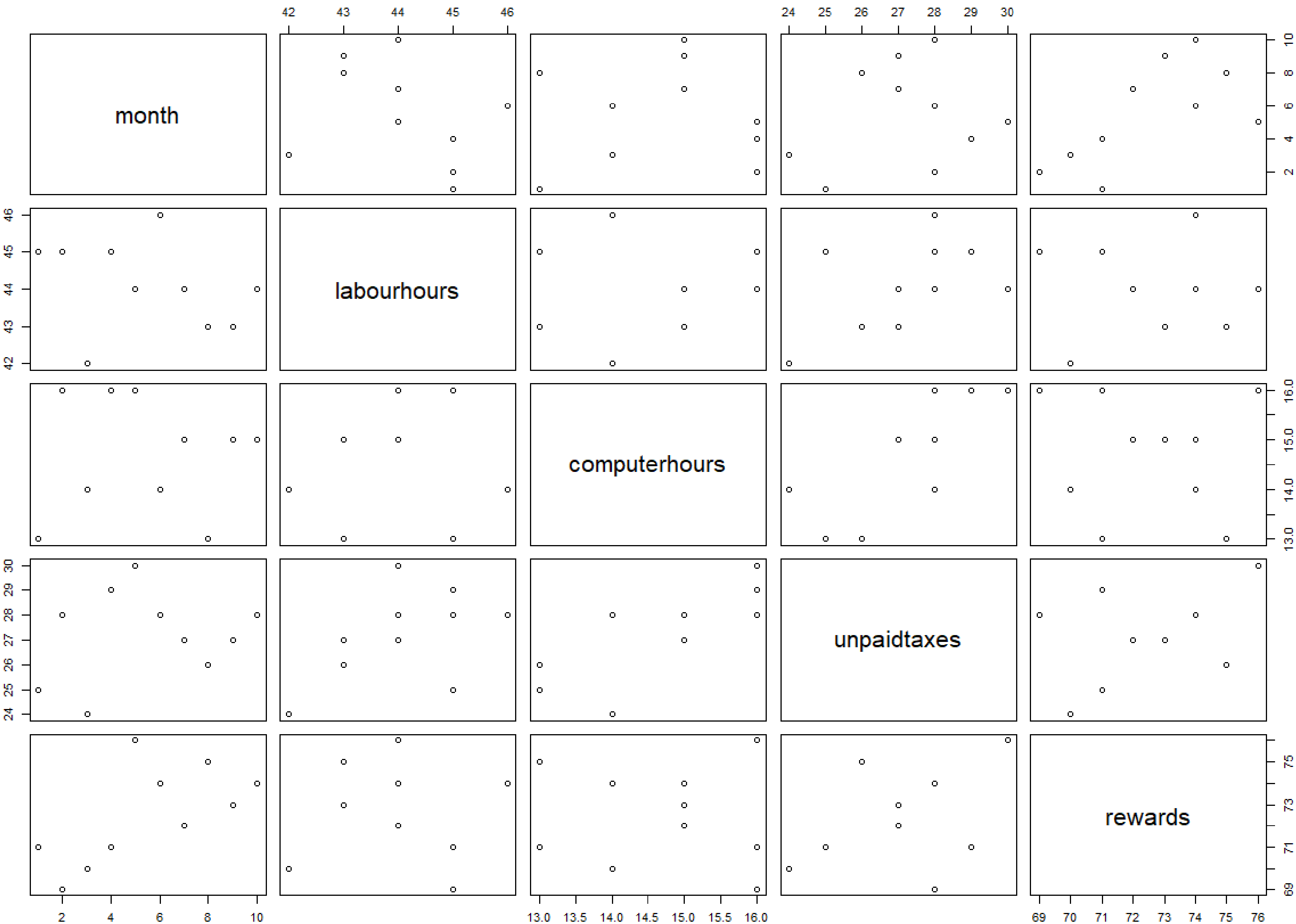
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Multiple Linear Regression: Example

- Example: Data for multiple Linear Regression

X	month	labourhours	computerhours	unpaidtaxes	rewards
1	jan	45	16	29	71
2	feb	42	14	24	70
3	mar	44	15	27	72
4	apr	45	13	25	71
5	may	43	13	26	75
6	jun	46	14	28	74
7	jul	44	16	30	76
8	aug	45	16	28	69
9	sep	44	15	28	74
10	oct	43	15	27	73

Data Visualization



Multiple Linear Regression Model

- Model-1
- Assume that unpaid taxes are dependent on
  - Labour hours
  - Computer hours

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.24668 -0.74702 -0.02321  0.51956  1.42706

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -13.8196    13.3233  -1.037  0.33411
labourhours     0.5637     0.3033   1.859  0.10543
computerhours   1.0995     0.3131   3.511  0.00984 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.071 on 7 degrees of freedom
Multiple R-squared:  0.7289,	Adjusted R-squared:  0.6515
F-statistic: 9.411 on 2 and 7 DF,  p-value: 0.01037
```

Multiple Linear Regression Model

- Model-2
- Assume unpaid taxes are also dependent on reqards
  - Labour hours
  - Computer hours
  - **Rewards**

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29080 -0.11604 -0.09998  0.09102  0.44452 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -45.79635    4.87765  -9.389 8.29e-05 ***
labourhours     0.59697    0.08112   7.359 0.000323 ***
computerhours   1.17684    0.08407  13.998 8.29e-06 ***
rewards         0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834,	Adjusted R-squared:  0.9751 
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

Comparing the models …

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.24668 -0.74702 -0.02321  0.51956  1.42706 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -13.8196    13.3233  -1.037  0.33411   
labourhours     0.5637     0.3033   1.859  0.10543   
computerhours   1.0995     0.3131   3.511  0.00984 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.071 on 7 degrees of freedom
Multiple R-squared:  0.7289,	Adjusted R-squared:  0.6515 
F-statistic: 9.411 on 2 and 7 DF,  p-value: 0.01037
```

MODEL-1

MODEL-2

Compare
- Residuals
- Residual Std. Error
- p-values
- R-squared
- F-Statistic & p-value

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29080 -0.11604 -0.09998  0.09102  0.44452 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -45.79635    4.87765  -9.389 8.29e-05 ***
labourhours     0.59697    0.08112   7.359 0.000323 ***
computerhours   1.17684    0.08407  13.998 8.29e-06 ***
rewards         0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834,	Adjusted R-squared:  0.9751 
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

Understanding Adjusted R-squared

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29080 -0.11604 -0.09998  0.09102  0.44452 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -45.79635    4.87765  -9.389 8.29e-05 ***
labourhours     0.59697    0.08112   7.359 0.000323 ***
computerhours   1.17684    0.08407  13.998 8.29e-06 ***
rewards         0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834,    Adjusted R-squared:  0.9751 
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

Adjusted R-squared

- Penalizes the addition of **insignificant** independent variables

Understanding F-Statistic and its p-value

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29080 -0.11604 -0.09998  0.09102  0.44452 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -45.79635    4.87765  -9.389 8.29e-05 ***
labourhours    0.59697    0.08112   7.359 0.000323 ***
computerhours  1.17684    0.08407  13.998 8.29e-06 ***
rewards        0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834,	Adjusted R-squared:  0.9751 
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

Larger the value of F-statistic (coupled with p-value less than 0.05), better the Regression

For the Multiple Linear Regression to be valid: $\beta_j \neq 0 \text{ for at least one } j$

Alternately, the regression is invalid if: $\beta_1 = \cdots = \beta_k = 0$

We need to check and ensure that second statement is not statistically true. This is done using the mechanism of statistical **Hypothesis Testing**. The following terms and ratios are calculated as part of this process:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$$

$$\boxed{F = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \sim F_{k,n-k-1}}$$

Understanding F-Statistic and its p-value

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29080 -0.11604 -0.09998  0.09102  0.44452 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -45.79635    4.87765  -9.389 8.29e-05 ***
labourhours    0.59697    0.08112   7.359 0.000323 ***
computerhours  1.17684    0.08407  13.998 8.29e-06 ***
rewards        0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834,	Adjusted R-squared:  0.9751 
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

Larger the value of F-statistic (coupled with p-value less than 0.05), better the Regression

$$F = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \sim F_{k,n-k-1}$$

ss_r = 29.10818

ss_{res} = 0.4912

F = $(ss_r / 3) / (ss_{res} / 6)$ = 118.50

Using the F-Distribution chart, it can be established that p-value corresponding to F-statistic $118.50_{(3,6)}$ = 9.935e-6

Comparing the models … F-Statistic

MODEL-1

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.24668 -0.74702 -0.02321  0.51956  1.42706 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -13.8196    13.3233  -1.037  0.33411   
labourhours     0.5637     0.3033   1.859  0.10543   
computerhours   1.0995     0.3131   3.511  0.00984 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.071 on 7 degrees of freedom
Multiple R-squared:  0.7289,	Adjusted R-squared:  0.6515 
F-statistic: 9.411 on 2 and 7 DF,  p-value: 0.01037
```

MODEL-2

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29080 -0.11604 -0.09998  0.09102  0.44452 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -45.79635    4.87765  -9.389 8.29e-05 ***
labourhours     0.59697    0.08112   7.359 0.000323 ***
computerhours   1.17684    0.08407  13.998 8.29e-06 ***
rewards         0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834,	Adjusted R-squared:  0.9751 
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

Model-2

- Larger F-Statistic with p-value almost zero
- Lower RSE
- Better R-Squared
- Higher confidence on individual coefficient estimates

Overall: A better model