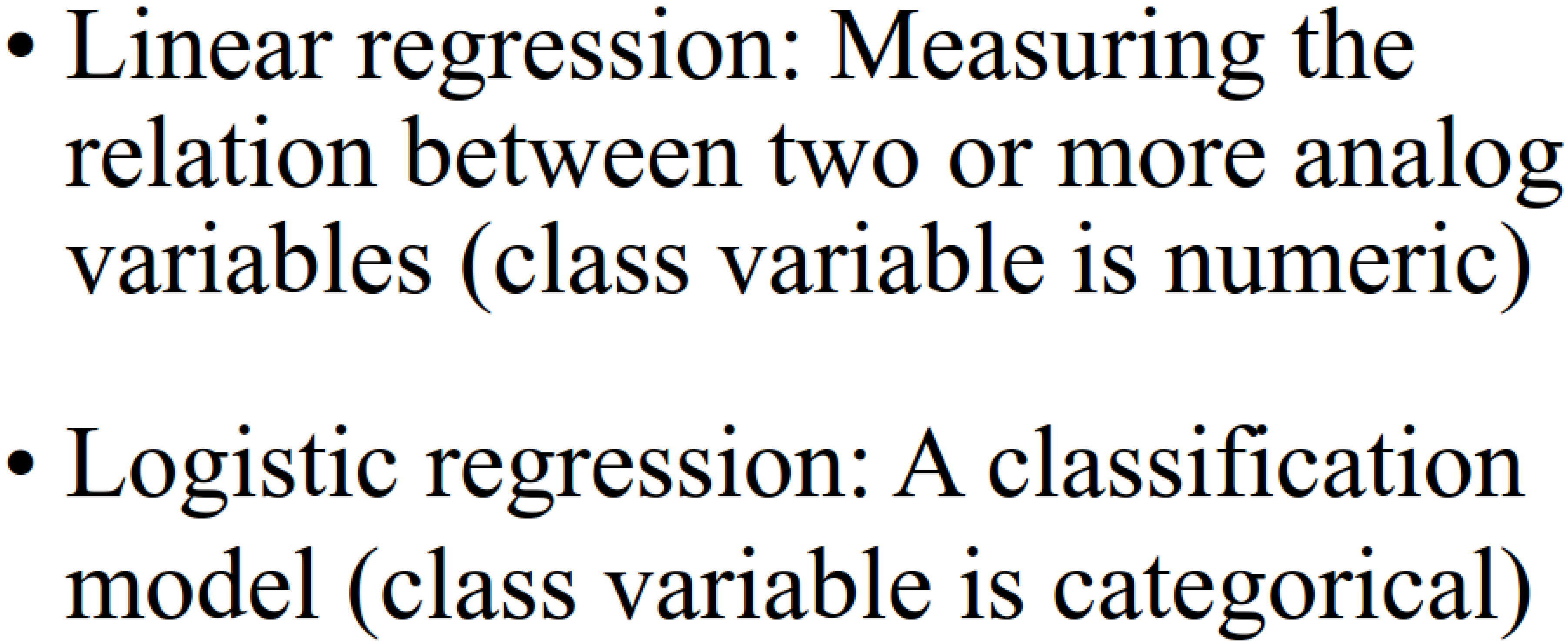
ML – Linear Regression.

Shah Ayub Quadri

Ayub.quadri89@gmail.com

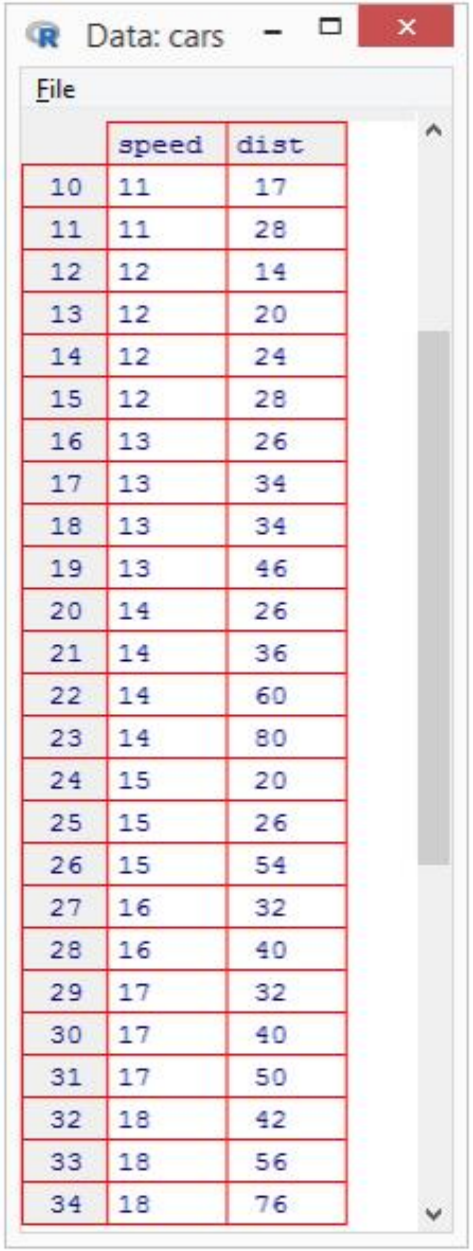
Supervised Learning



- Linear regression: Measuring the relation between two or more analog variables (class variable is numeric)

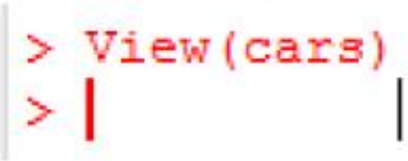
- Logistic regression: A classification model (class variable is categorical)

Speed vs Stopping distance



	speed	dist
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
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

The “cars” dataset in R contains 50 pairs of datapoints
for Speed(mph) vs stopping distance(ft), that were collected in 1920

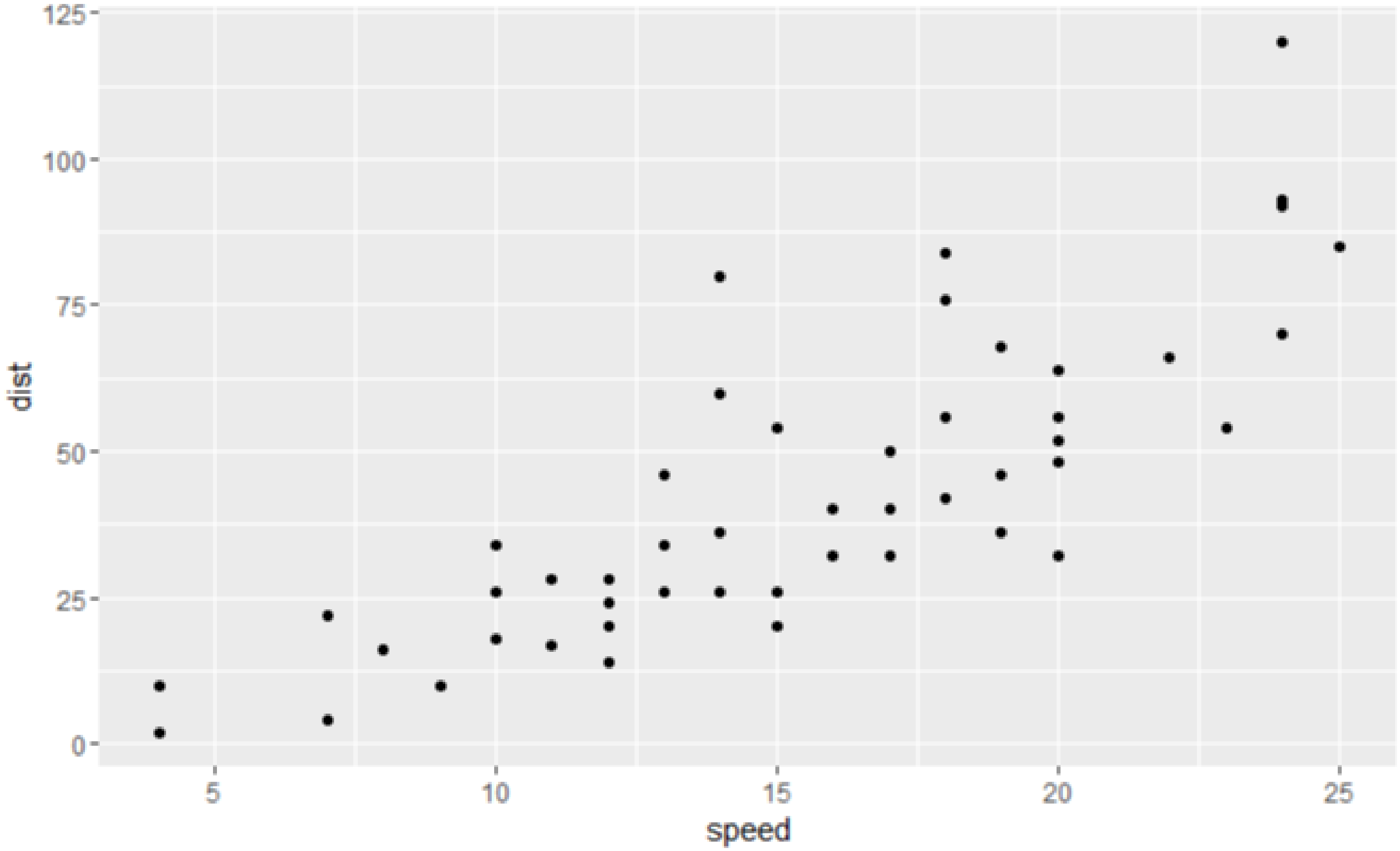


```
> View(cars)
>
```

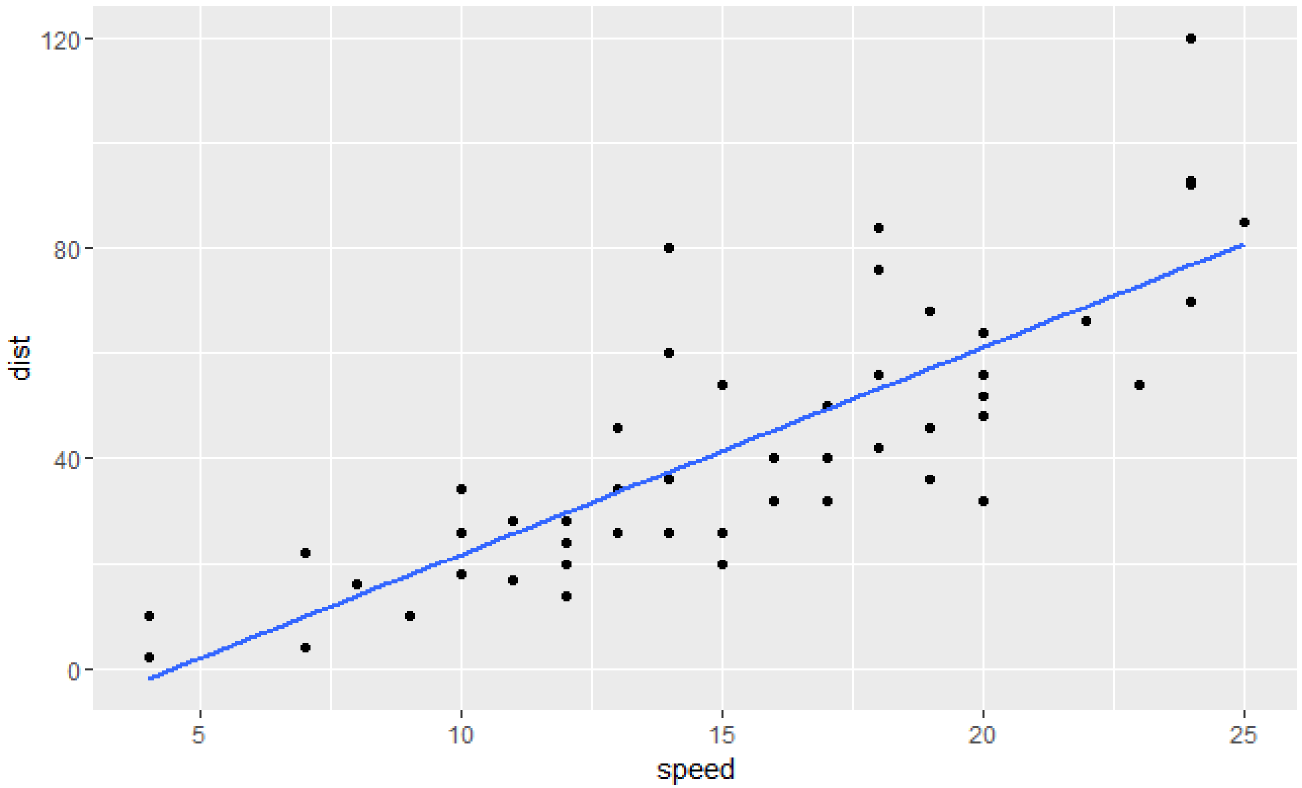
See: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html

Independent variable (explanatory) – Speed (mph) – Plotted on X-axis

Dependent variable (response) – Stopping distance(ft) – Plotted on Y-axis



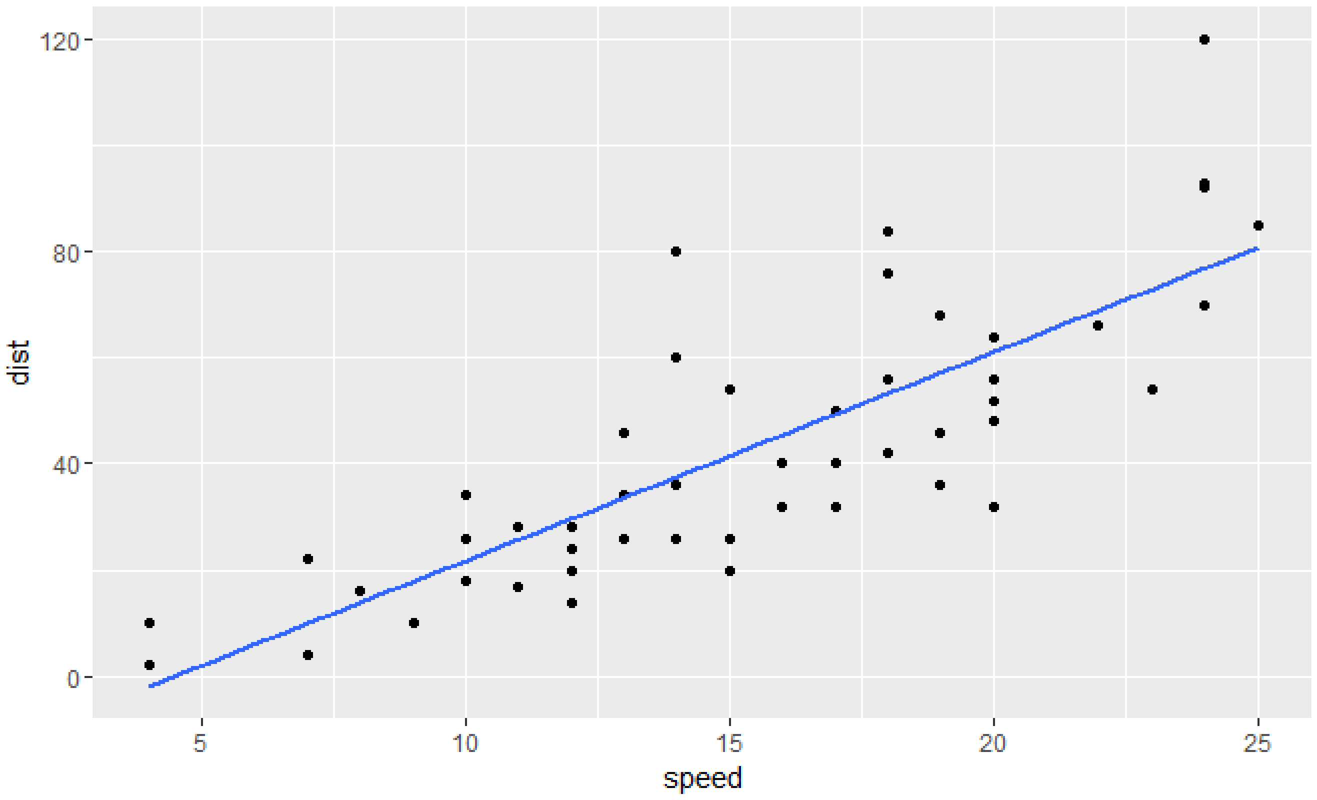
Speed vs Stopping distance



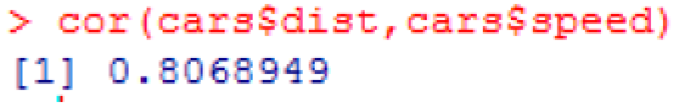
$$y = 3.93\,x - 17.58$$

```
> lmcars <- lm(dist~speed, data=cars)
> summary(lmcars)
```

Correlation: Speed vs Stopping distance

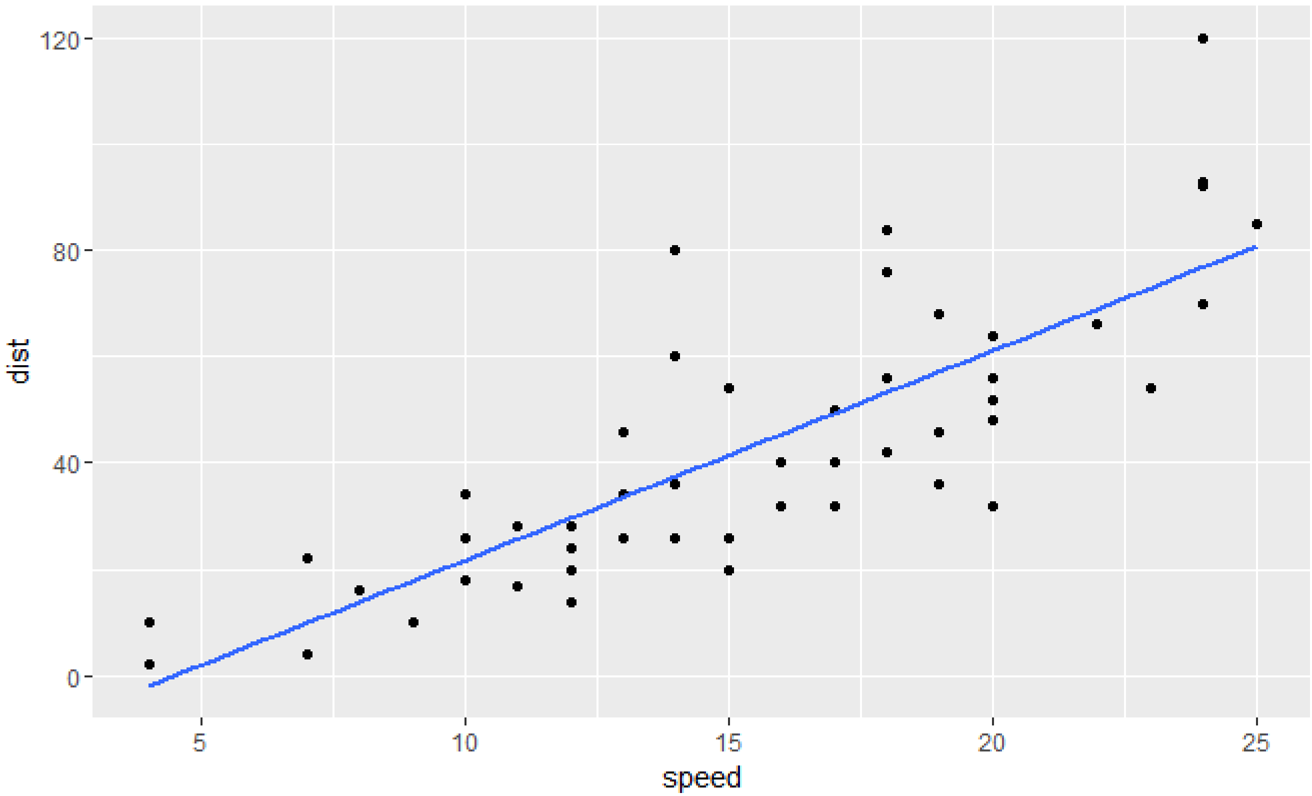


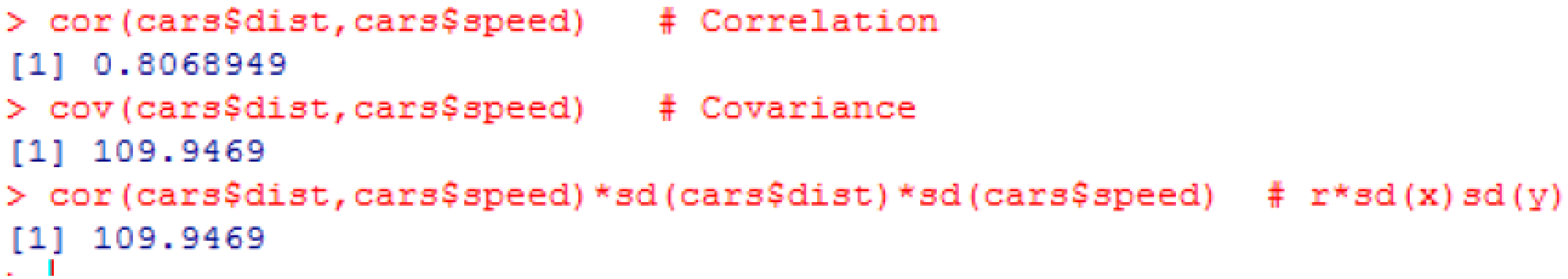
$$y = 3.93\,x - 17.58$$



```
> cor(cars$dist,cars$speed)
[1] 0.8068949
```

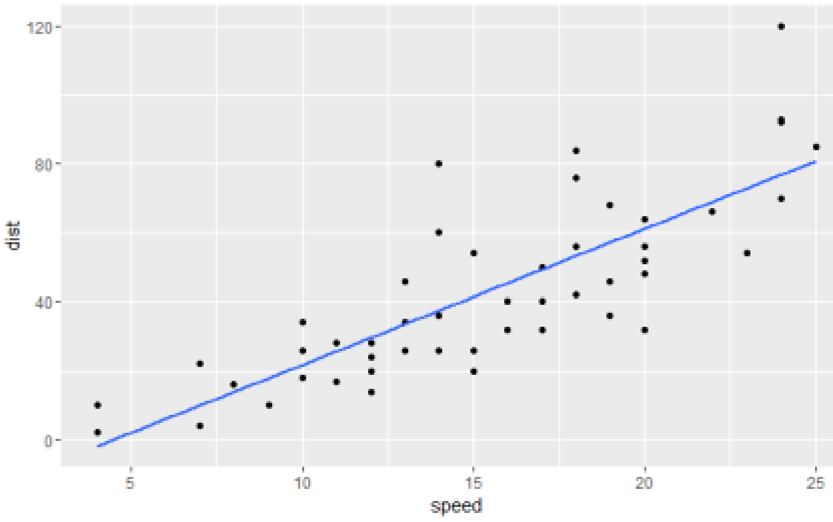
Covariance: Speed vs Stopping distance

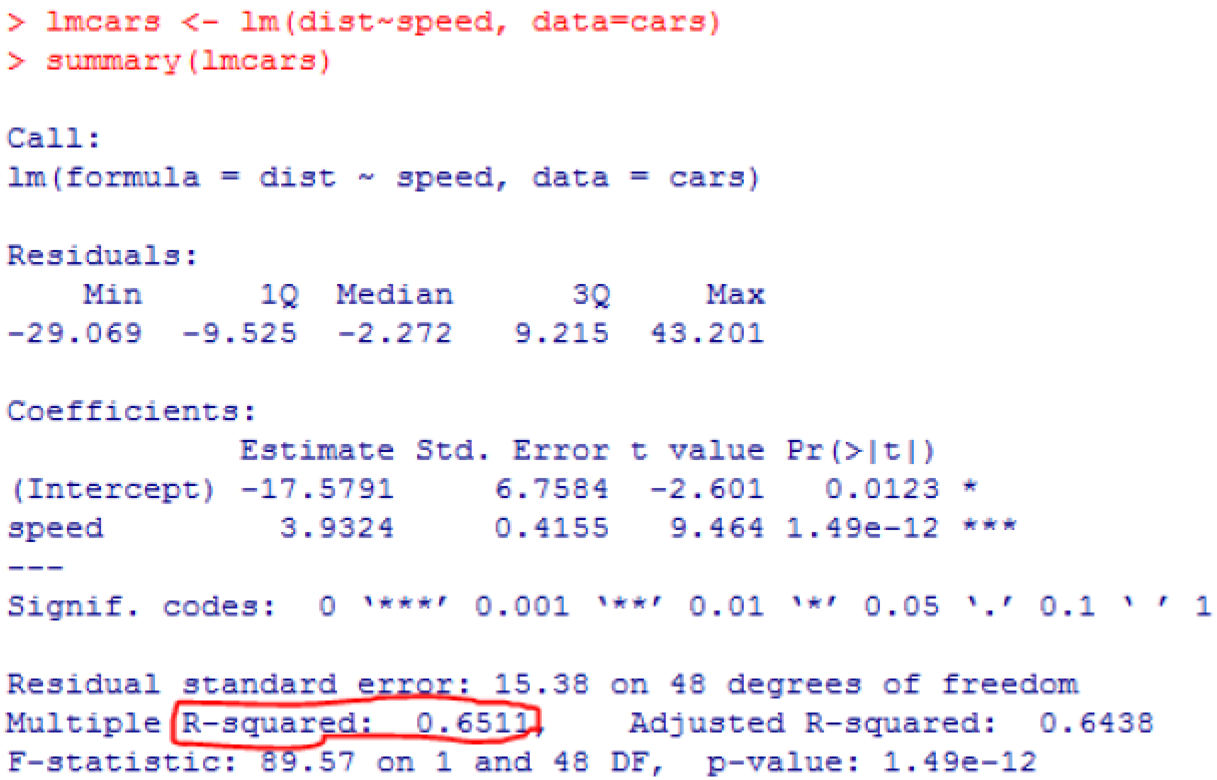




```
> cor(cars$dist,cars$speed)    # Correlation
[1] 0.8068949
> cov(cars$dist,cars$speed)    # Covariance
[1] 109.9469
> cor(cars$dist,cars$speed)*sd(cars$dist)*sd(cars$speed)   # r*sd(x)sd(y)
[1] 109.9469
```

R^2: Speed vs Stopping distance

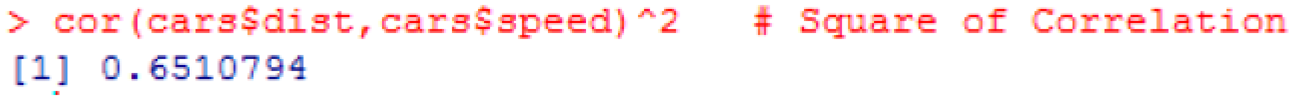




```
> lmcars <- lm(dist~speed, data=cars)
> summary(lmcars)

Call:
lm(formula = dist ~ speed, data = cars)

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
```

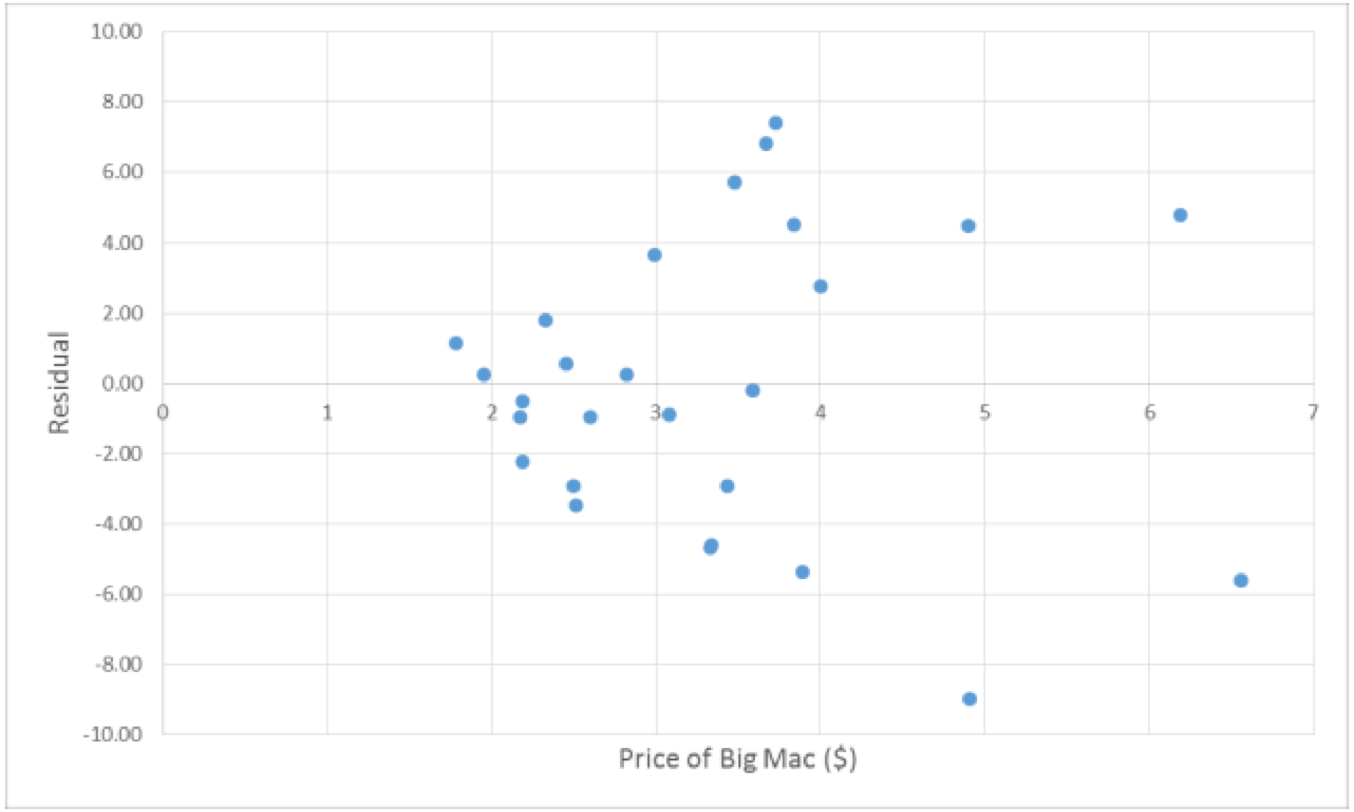
R^2 suggests that 65% of variation in 'y' can be explained by variation in 'x'

$R^2 = 0.6511$



```
> cor(cars$dist,cars$speed)^2   # Square of Correlation
[1] 0.6510794
```

Residual Analysis



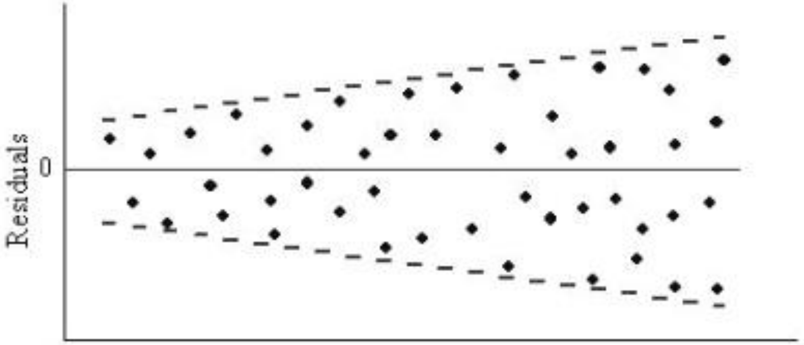
Can be used to locate outliers.

Residual Analysis: https://www.stat.berkeley.edu/~stark/SticiGui/Text/regressionDiagnostics.htm

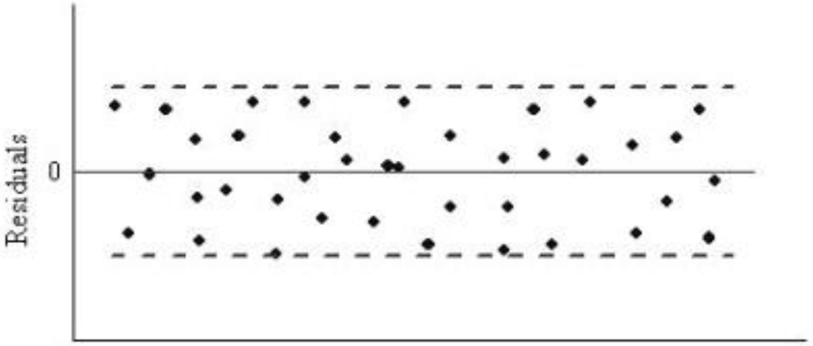
Assumptions of the Regression Model

- The error terms have constant variances (homoscedasticity as opposed to heteroscedasticity)

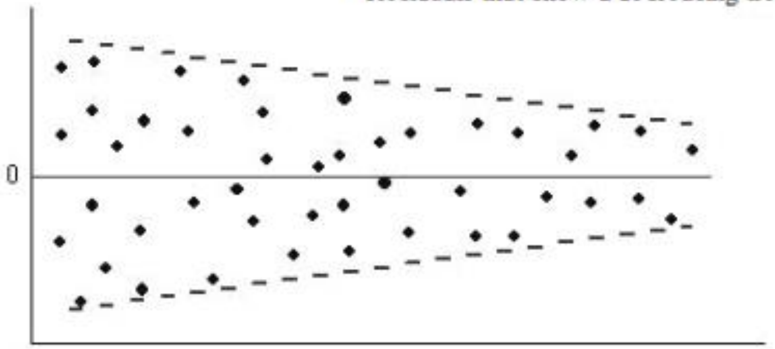
Residuals that show an increasing trend



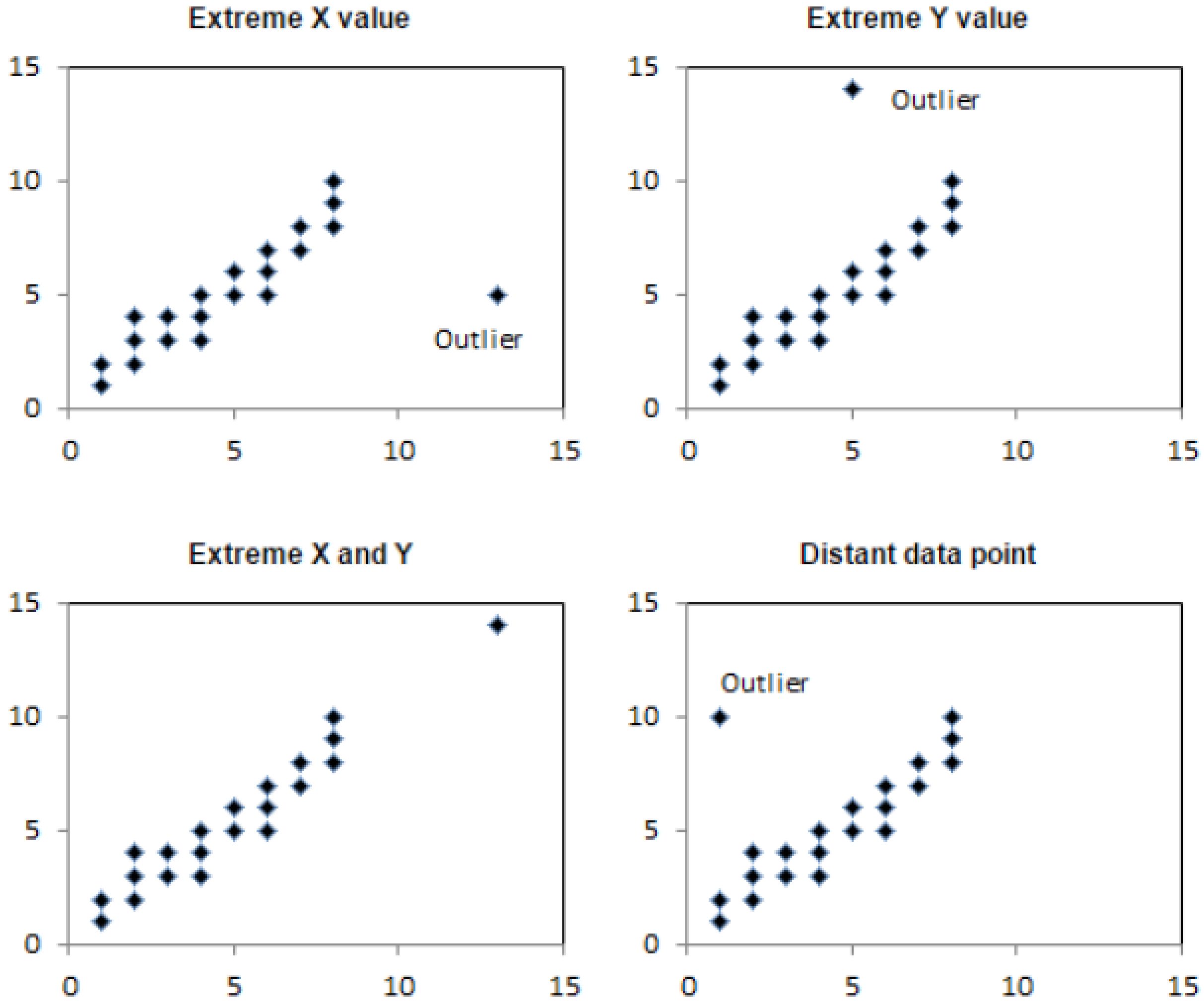
Constant variance



Residuals that show a decreasing trend



Outliers



Extreme X value

Extreme Y value

Extreme X and Y

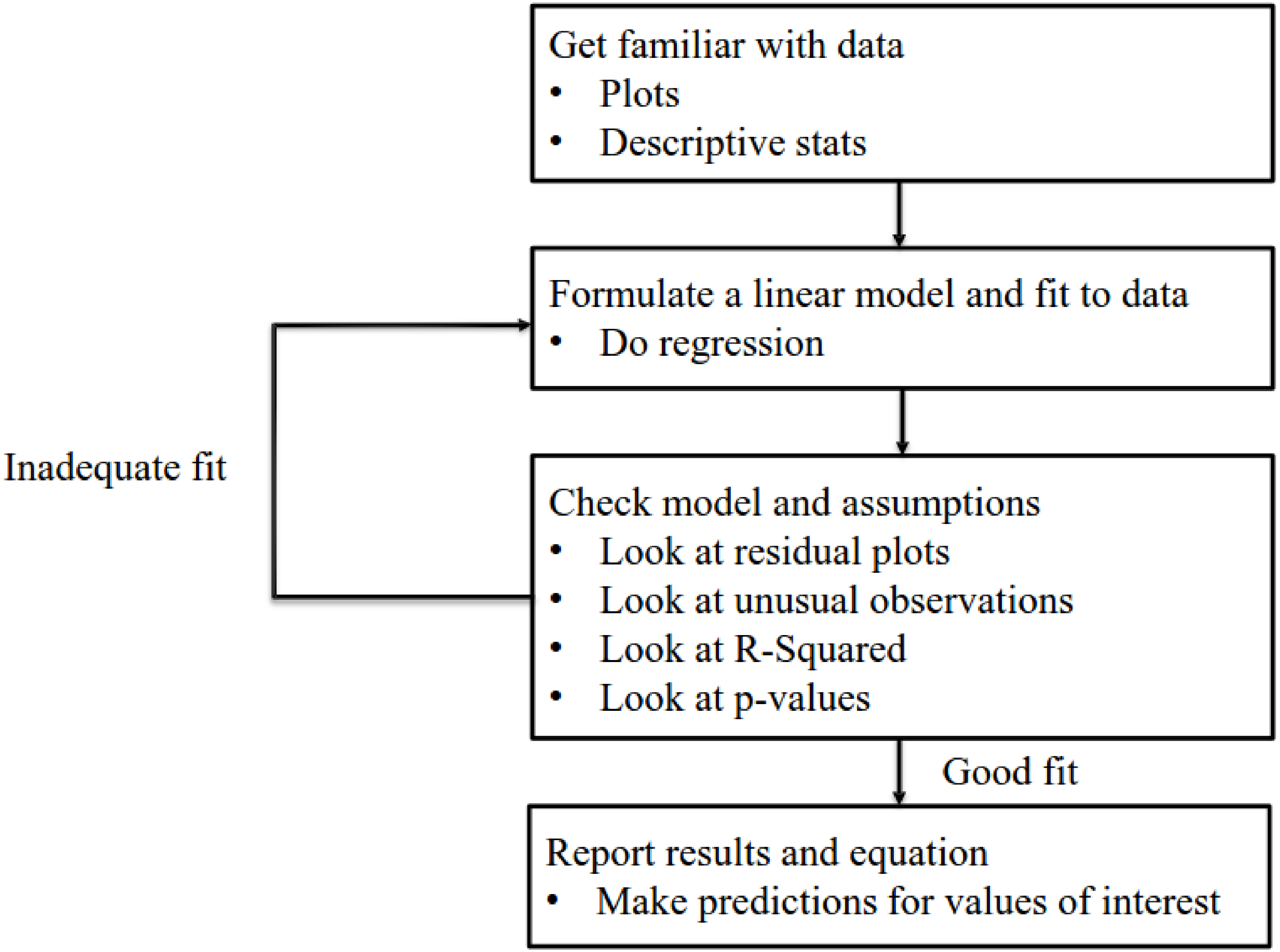
Distant data point

Outliers do not follow the general trend of the rest of the data

Outliers typically have a large residual.

Source: http://stattrek.com/regression/influential-points.aspx?Tutorial=AP

Simple Linear Regression - Steps



Get familiar with data

- Plots
- Descriptive stats

Formulate a linear model and fit to data

- Do regression

Check model and assumptions

- Look at residual plots
- Look at unusual observations
- Look at R-Squared
- Look at p-values

Inadequate fit

Good fit

Report results and equation

- Make predictions for values of interest

Multiple Linear Regression

- Linear regression models the effect of one independent variable, x, on one dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_n x_n + \varepsilon$$

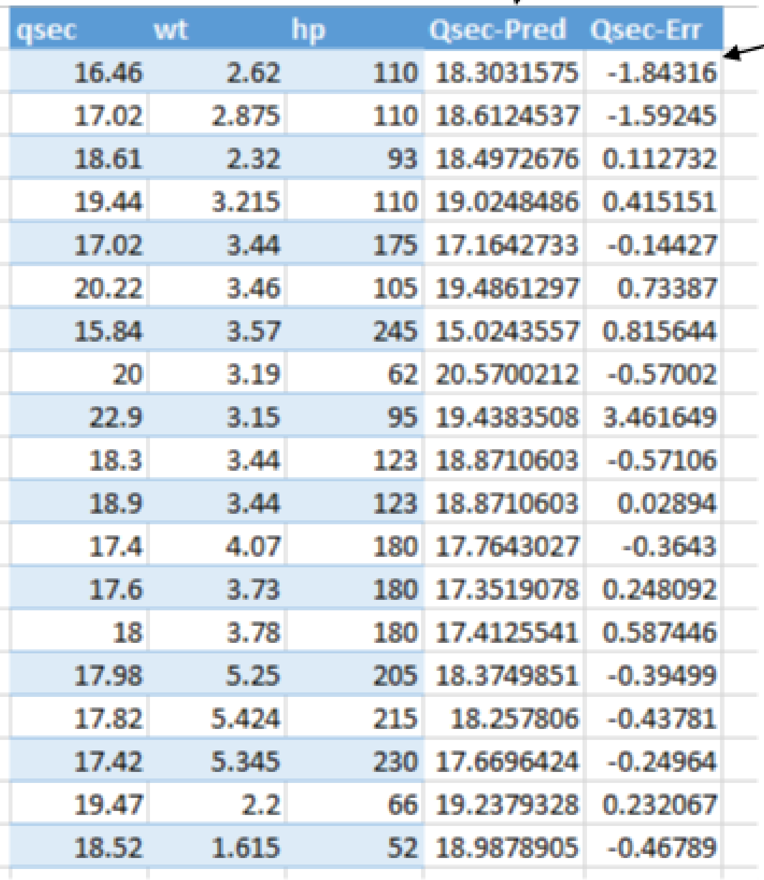
- The β parameters reflect the **independent contribution** of each independent variable, x, to the value of the dependent variable, y.

Cars Dataset (MTcars)

model	mpg	wt	hp	qsec
Mazda RX4	21	2.62	110	16.46
Mazda RX4 Wag	21	2.875	110	17.02
Datsun 710	22.8	2.32	93	18.61
Hornet 4 Drive	21.4	3.215	110	19.44
Datsun 710	22.8	2.32	93	18.61
Hornet 4 Drive Valiant	21.4 18.1	3.215 3.46	110 105	19.44 20.22
Duster 360	14.3	3.57	245	15.84
Merc 240D	24.4	3.19	62	20
Merc 230	22.8	3.15	95	22.9
Merc 280	19.2	3.44	123	18.3
Merc 280C	17.8	3.44	123	18.9
Merc 450SE	16.4	4.07	180	17.4
Merc 450SL	17.3	3.73	180	17.6
Merc 450SLC	15.2	3.78	180	18
Cadillac Fleetwood	10.4	5.25	205	17.98
Lincoln Continental	10.4	5.424	215	17.82
Chrysler Imperial	14.7	5.345	230	17.42
Fiat 128	32.4	2.2	66	19.47
Honda Civic	30.4	1.615	52	18.52

Mpg=Miles/gallon
Wt = weight
Hp = horsepower
Qsec=time to go cover a quarter mile from start

Qsec predicted from (wt,hp)



qsec	wt	hp	Qsec-Pred	Qsec-Err
16.46	2.62	110	18.3031575	-1.84316
17.02	2.875	110	18.6124537	-1.59245
18.61	2.32	93	18.4972676	0.112732
19.44	3.215	110	19.0248486	0.415151
17.02	3.44	175	17.1642733	-0.14427
20.22	3.46	105	19.4861297	0.73387
15.84	3.57	245	15.0243557	0.815644
20	3.19	62	20.5700212	-0.57002
22.9	3.15	95	19.4383508	3.461649
18.3	3.44	123	18.8710603	-0.57106
18.9	3.44	123	18.8710603	0.02894
17.4	4.07	180	17.7643027	-0.3643
17.6	3.73	180	17.3519078	0.248092
18	3.78	180	17.4125541	0.587446
17.98	5.25	205	18.3749851	-0.39499
17.82	5.424	215	18.257806	-0.43781
17.42	5.345	230	17.6696424	-0.24964
19.47	2.2	66	19.2379328	0.232067
18.52	1.615	52	18.9878905	-0.46789

The part of Qsec unexplained by (wt,hp)

References

Residual Analysis

https://www.stat.berkeley.edu/~stark/SticiGui/Text/regressionDiagnostics.htm

Outliers

http://stattrek.com/regression/influential-points.aspx?Tutorial=AP

Homoscedasticity

https://www.youtube.com/watch?v=Yf1efX-2LXI