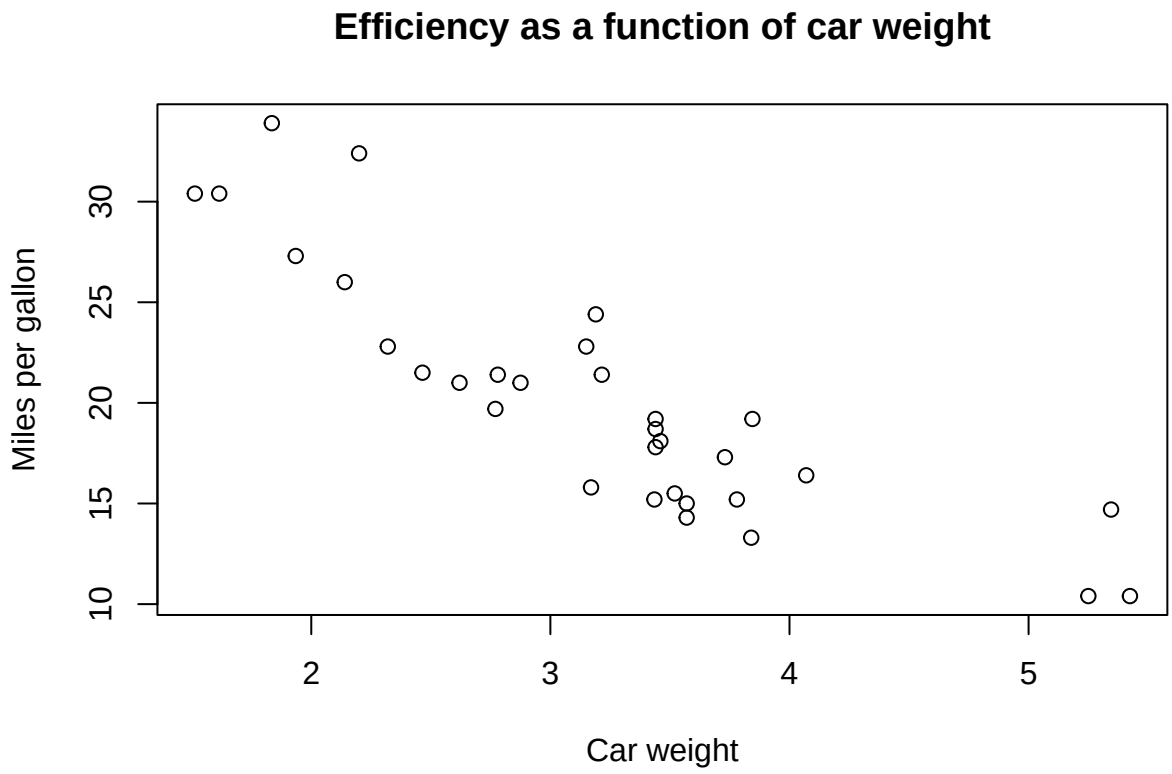
Linear Regression

December 10, 2025

Contents

1 Scatter plot example

```
plot(y = mtcars$mpg, x = mtcars$wt,
     main = "Efficiency as a function of car weight",
     ylab = "Miles per gallon", xlab = "Car weight")
```




```
# alternative: plot(mpg~wt, data = mtcars)
```

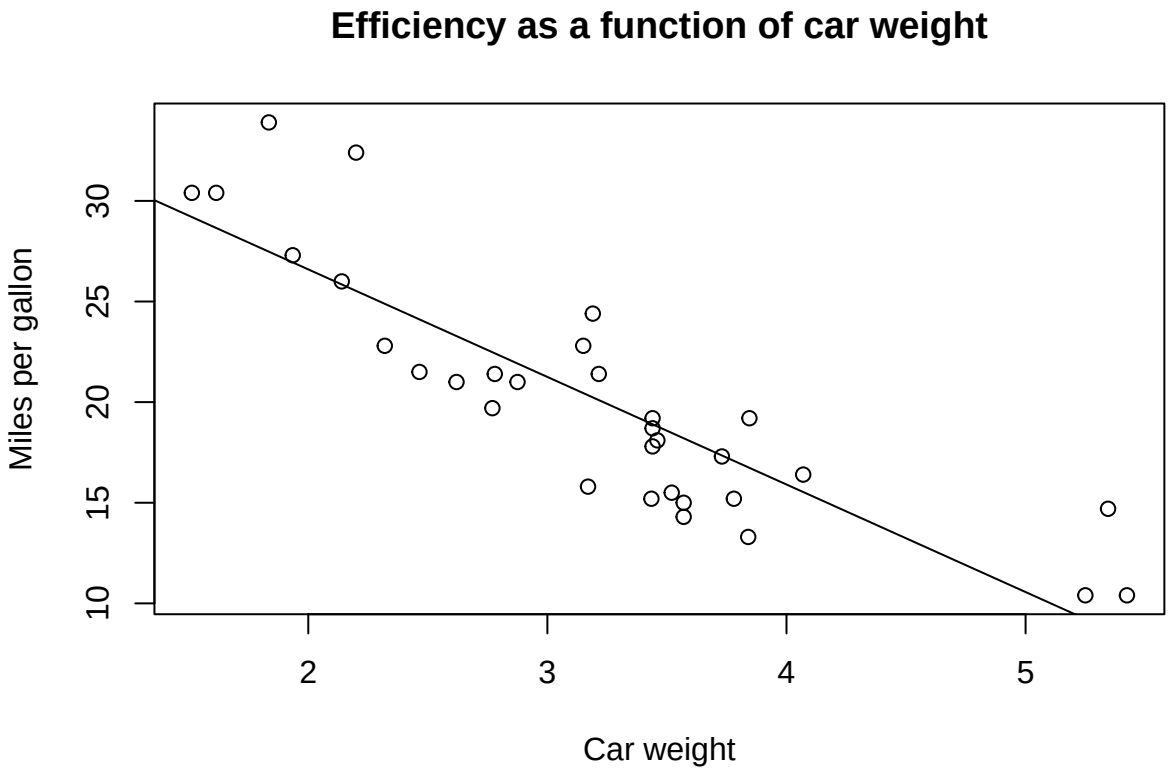
2 Linear model example

```
fit <- lm(formula = mtcars$mpg ~ mtcars$wt)

# alternative: lm(formula = mpg ~ wt, data = mtcars)

plot(y = mtcars$mpg, x = mtcars$wt,
     main = "Efficiency as a function of car weight",
     ylab = "Miles per gallon", xlab = "Car weight")
```

```r
# Show fitted line
abline(fit)
```



3 Measuring the strength of association

$$H_0 : B = 0 \quad \text{(Slope of regression line is zero)}$$

```r
# Get the p-value using `summary()`
summary(fit)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## mtcars$wt    -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
# To get the p-value of the slope

# This displays the slope and intercept
# along with their p-values 'Pr(>|t|)'
# By default the p-values are two-sided
summary(fit)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.285126   1.877627 19.857575 8.241799e-19
## mtcars$wt   -5.344472   0.559101 -9.559044 1.293959e-10
```

```
# To get goodness of fit
# The model explains '100*r^2' percent of the variance in y
summary(fit)$r.squared
```

```
## [1] 0.7528328
```

4 Multiple Regression

```
# Use '+' to add more independent variables (x1,x2,...,xn)
fit_multiple <- lm(mpg~wt+disp,data=mtcars)
summary(fit_multiple)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.96055404 2.164539504 16.151497 4.910746e-16
## wt          -3.35082533 1.164128079 -2.878399 7.430725e-03
## disp        -0.01772474 0.009190429 -1.928609 6.361981e-02
```

5 ANOVA

```
# Used when comparing means of multiple categories.
# Used for analyzing the results of an experiment with multiple groups.

# Example

light <-
data.frame(experiment=
letters[c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,5,5,5,
5,5,5,5,5,5,5)],speed=c(85,74,90,107,93,85,95,98,98,88,100,98,93,
65,76,81,100,100,96,96,96,94,96,94,88,80,85,88,90,84,83,79,81,88,
88,83,80,79,76,80,88,88,88,86,72,72,62,86,97,95,88,91,85,87,84,84,
85,84,84,84,89,81,81,82,80,77,76,74,75,76,91,92,89,86,88,72,84,85
,85,78,89,84,78,81,76,81,79,81,82,85,87,87,81,74,81,94,95,80,81,87))

head(light)
```

```
##   experiment speed
## 1          a    85
## 2          a    74
## 3          a    90
## 4          a   107
## 5          a    93
## 6          a    85
```

```
# Null: mean_a = mean_b = mean_c = mean_d = mean_e
# To get the p-value
# Since 0.00311 < 0.05 we reject the null
summary(aov(speed~experiment,data=light))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## experiment   4    945  236.29   4.288 0.00311 **
## Residuals   95   5235   55.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```