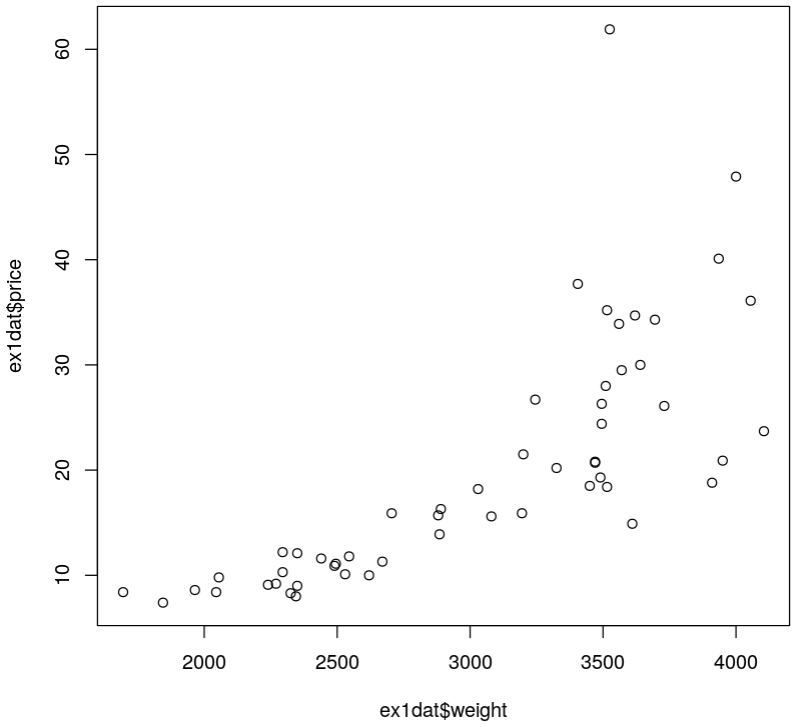
In [3]:
```
# Read data file "cars.csv" from openintro.org site
#
ex1dat = read.csv(file="https://www.openintro.org/data/csv/cars93.csv", header=TRUE, sep=",")
head(ex1dat)
#lmresults = lm(write ~ socst+math+read+science, data=ex1dat)
#summary (lmresults)
```

A data.frame: 6 × 6

	type	price	mpg_city	drive_train	passengers	weight
	<fct>	<dbl>	<int>	<fct>	<int>	<int>
1	small	15.9	25	front	5	2705
2	midsize	33.9	18	front	5	3560
3	midsize	37.7	19	front	6	3405
4	midsize	30.0	22	rear	4	3640
5	midsize	15.7	22	front	6	2880
6	large	20.8	19	front	6	3470

In [7]:
```
# Explore relationship between price and other variables.
# Let's consider weight first
#
plot(ex1dat$weight, ex1dat$price)
cat("The correlation =", cor(ex1dat$weight, ex1dat$price))
```

```
The correlation = 0.758112
```



In [10]:
```
# Fit linear regression to model price vs weight
#
lmresults = lm (price ~ weight, data=ex1dat)
summary (lmresults)
```

```
Call:
lm(formula = price ~ weight, data = ex1dat)

Residuals:
    Min      1Q  Median      3Q     Max
-12.767  -3.766  -1.155   2.568  35.440

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.295205   4.915159  -4.129 0.000132 ***
weight        0.013264   0.001582   8.383 3.17e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.575 on 52 degrees of freedom
Multiple R-squared:  0.5747,    Adjusted R-squared:  0.5666
F-statistic: 70.28 on 1 and 52 DF,  p-value: 3.173e-11
```

Exercise for students:

- Write the equation of the regression line.
- Interpret the slope.
- Interpret R^2.

Exercise 2 for students:

- Do a similar study for the relationship between price and mpg_city.
- Write the equation of the regression line.
- Interpret the slope.
- Interpret R^2.

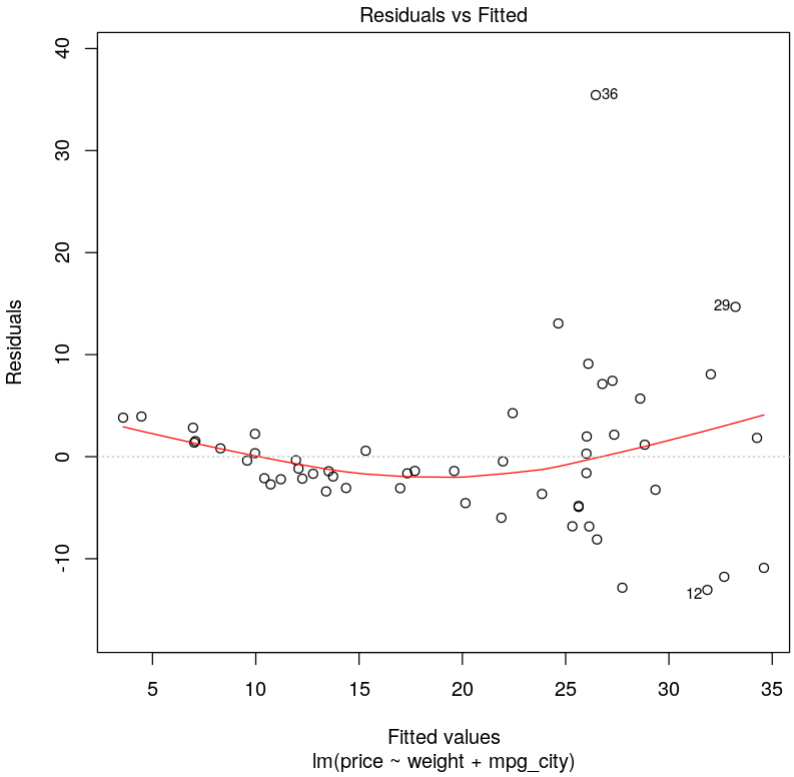
In [17]:
```
# Now let us explore the relationship between price and
# both  weight and mpg_city taken together.
#
lmresults = lm(price ~ weight+mpg_city, data=ex1dat)
summary (lmresults)
```

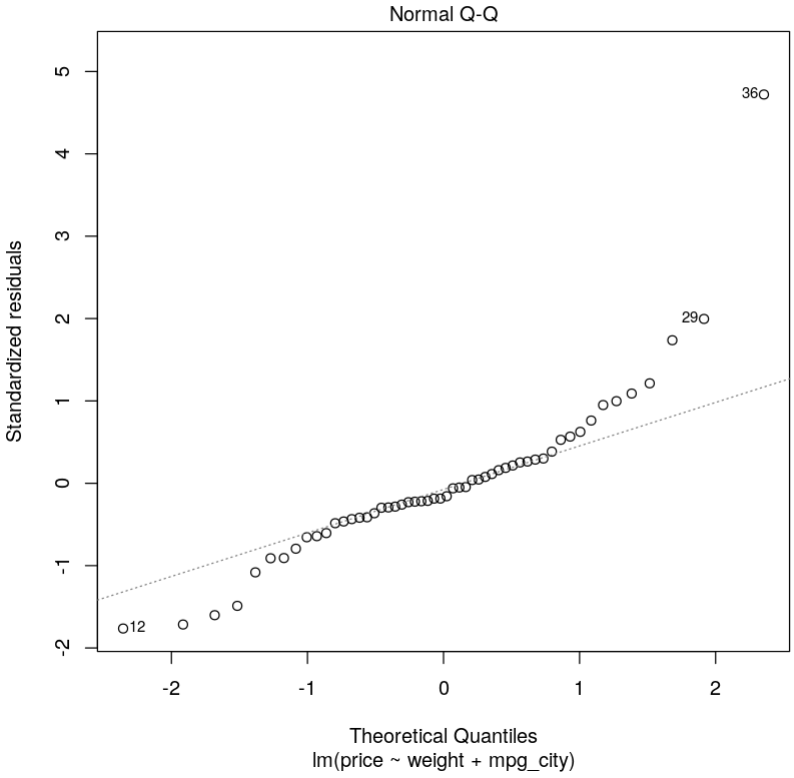
```
Call:
lm(formula = price ~ weight + mpg_city, data = ex1dat)

Residuals:
    Min      1Q  Median      3Q     Max
-13.059  -3.209  -1.284   2.108  35.442

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.836215  17.207332  -1.792   0.0791 .
weight        0.015121   0.003311   4.567 3.16e-05 ***
mpg_city      0.210219   0.328699   0.640   0.5253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.619 on 51 degrees of freedom
Multiple R-squared:  0.5781,    Adjusted R-squared:  0.5616
F-statistic: 34.94 on 2 and 51 DF,  p-value: 2.769e-10
```

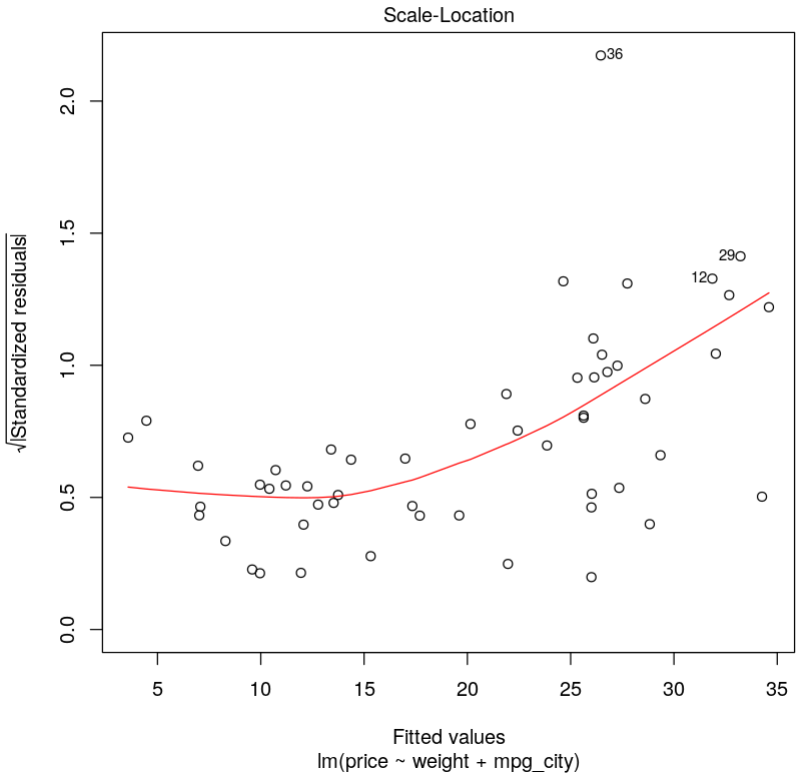
Exercise 3

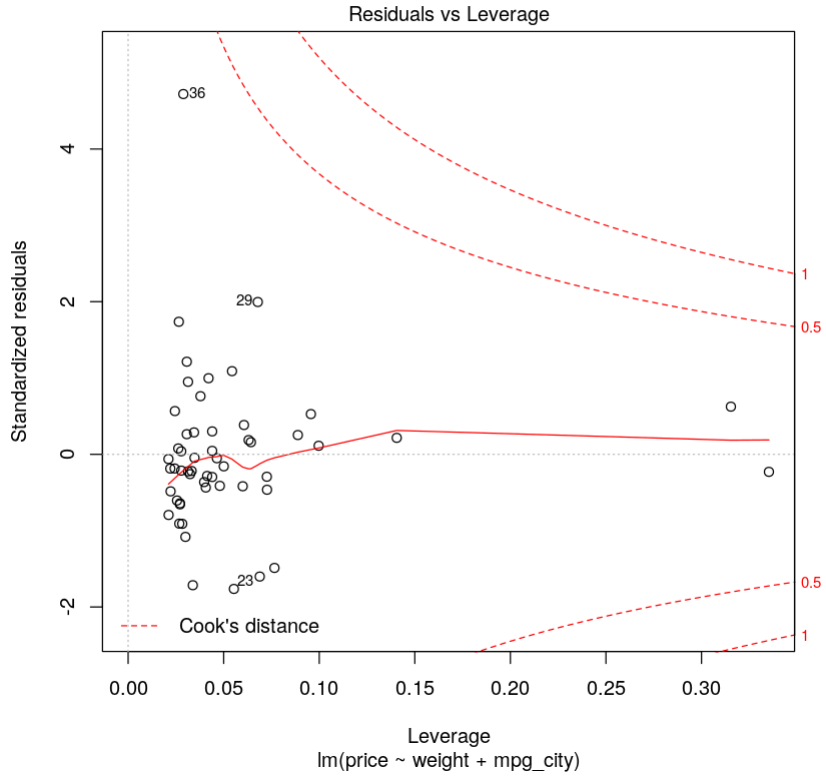
- Write the equation of the regression line.
- Interpret each slope.
- Interpret R^2.

In [20]:
```
# Various diagnostic plots based on residuals:
plot (lmresults)
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ weight + mpg_city)
36
29
12
40
30
20
10
0
-10
5
10
15
20
25
30
35


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(price ~ weight + mpg_city)
36
29
12
5
4
3
2
1
0
-1
-2


Scale-Location
√|Standardized residuals|
Fitted values
lm(price ~ weight + mpg_city)
36
29
12


Residuals vs Leverage
Standardized residuals
Leverage
lm(price ~ weight + mpg_city)
Cook's distance
36
29
23
1
0.5
0.5
1

Another dataset

Now let's try a dataset in which the response variable has stronger correlation with predictor variables. It contains 200 observations from the High School and Beyond survey conducted by the National Center for Educational Statistics.

In [21]:
```
ex2dat = read.csv(file="https://www.openintro.org/data/csv/hsb2.csv", header=TRUE, sep=",")
head(ex2dat)
```

A data.frame: 6 × 11

	id	gender	race	ses	schtyp	prog	read	write	math	science	socst
	<int>	<fct>	<fct>	<fct>	<fct>	<fct>	<int>	<int>	<int>	<int>	<int>
1	70	male	white	low	public	general	57	52	41	47	57
2	121	female	white	middle	public	vocational	68	59	53	63	61
3	86	male	white	high	public	general	44	33	54	58	31
4	141	male	white	high	public	vocational	63	44	47	53	56
5	172	male	white	middle	public	academic	47	52	57	53	61
6	113	male	white	middle	public	academic	44	52	51	63	61

Exercise

- Make pairwise scatterplots of all the test scores against each other.
- Find pairwise correlation of all the test scores with each other.
- Create a linear model to predict "read" scores from various other predictors. Which predictor works best?
- Create a MLR model to predict "read" scores.
- Make plots showing residuals diagnostics.
- Determine whether the conditions for MLR are met.

In [ ]:
```

```