Intermediate Econometrics

5th/7th November 2025 - Vincenzo Gioia

Linear Model: Example

Advertising Data Set

- Let's suppose we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product
- The **Advertising** data set consists of the **sales** of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: **TV**, **Radio**, and **Newspaper**

First question?

- Identify the outcome variable and the predictors

First Answer

- Outcome: **Sales** (y)
- Predictors: **TV** (x_2), **Radio** (x_3), and **Newspaper** (x_4)

Linear Model: Example

Advertising Data Set

- It is not possible for our client to directly increase sales of the product …
- … but they can control the advertising expenditure in each of the three media

Goal

- If we determine that there is an association between advertising and sales, we can instruct our client to adjust the advertising budgets, thereby indirectly increasing the sales
- **Develop a model that can be used to predict sales on the basis of the three media budgets**

Linear Model: Example

Reading the Data

- The data are stored in a .csv file (available on the moodle page)
- The R function to read a .csv file is **read.csv()**
- In this case, no further arguments are passed to the function (see the help)
- **Exercise: Try to read the data using read.csv2() function**

```
1 Advertising <- read.csv("Advertising.csv")
```

The working directory

- When you run an R code requiring external sources (in this case the data), you must set the correct working directory. This can be done:

1. Using the graphical interface: move the pointer over the R file, right-click it, and select *Set Working Directory*
2. By hand: setting the working directory by using **setwd()**

Linear Model: Example

Explore the Structure of the data set

- After verifying that the process of reading the data has been performed successfully, we can analyze the data structure (dimension of the data set, name of variables, type of the variables)
- We can do that using the functions **dim()**, **names()** and **str()**

```
 1  dim(Advertising)
```

```
[1] 200   5
```

```
 1  names(Advertising)
```

```
[1] "X"         "TV"        "Radio"     "Newspaper" "Sales"
```

```
 1  str(Advertising)
```

```
'data.frame':   200 obs. of  5 variables:
 $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ TV        : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio     : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales     : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

Linear Model: Example

Univariate Exploratory Analysis

- Having graphical (or numerical) information of the variables under analysis. This can be done

1. by using the **summary()** function
2. by using graphical tools

- In our example all the variables are continuos (we can use **hist()** for 2.)
- The sales are in thousands of units and the advertising budgets are in thousands of dollars

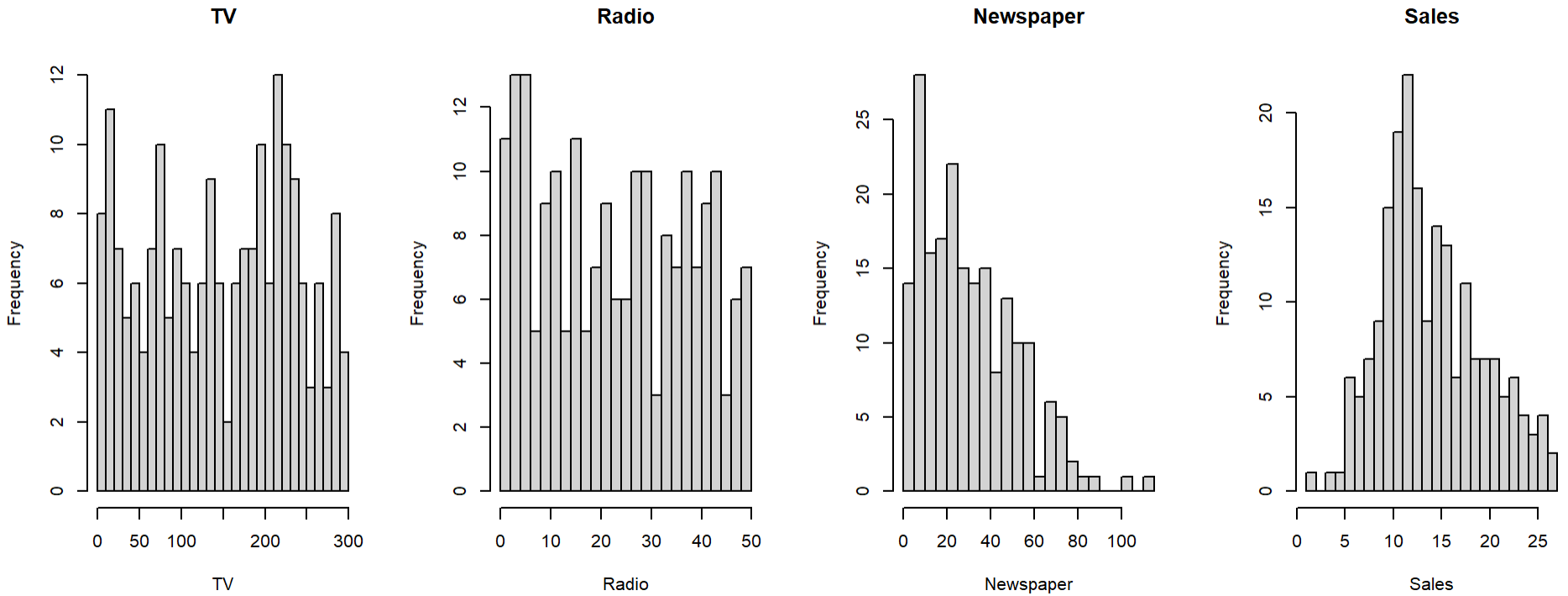
```
1  summary(Advertising[,-1])
```

```
       TV              Radio          Newspaper          Sales      
 Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60  
 1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38  
 Median :149.75   Median :22.900   Median : 25.75   Median :12.90  
 Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02  
 3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40  
 Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00  
```

Linear Model: Example

> **Univariate Exploratory Analysis**
>
> - Histograms

```
par(mfrow=c(1,4))
for (i in 2:5) hist(Advertising[,i], breaks = 30, main = names(Advertising)[i],
                    xlab = names(Advertising)[i])
```


TV
Frequency
0
2
4
6
8
10
12
0
50
100
200
300
TV
Radio
Frequency
0
2
4
6
8
10
12
0
10
20
30
40
50
Radio
Newspaper
Frequency
0
5
10
15
20
25
0
20
40
60
80
100
Newspaper
Sales
Frequency
0
5
10
15
20
0
5
10
15
20
25
Sales

Linear Model: Example

Problems and questions

- We are asked to suggest, on the basis of the data, a marketing plan that will result in high product sales
- What information would be useful in order to provide such a recommendation?

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media are associated with sales?
4. How large is the association between each medium and sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there any synergy among the advertising media?

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Our first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales. If the evidence is weak, one might argue that no money should be spent on advertising
- We could start our exploration by fitting simple linear regressions, each of which uses a different advertising medium as a predictor
- Let's consider a simple linear regression of sales on TV (x_2):

$$Y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, for $i = 1, \ldots, n$, indepedently

```
1  fitTV <- lm(Sales ~ TV, data = Advertising)
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- By regressing sales on TV we get the estimated regression line

$$y = 7.033 + 0.047x_2$$

1. $\hat{\beta}_1 = 7.033$: For $0\$$ spent in TV the average number of sales is 7.033 units
2. $\hat{\beta}_2 = 0.047$: A $1000\$$ increasing in spending on TV advertising is associated with an increase of sales of around 47.5 units on average

```
summary(fitTV)$coefficients
```

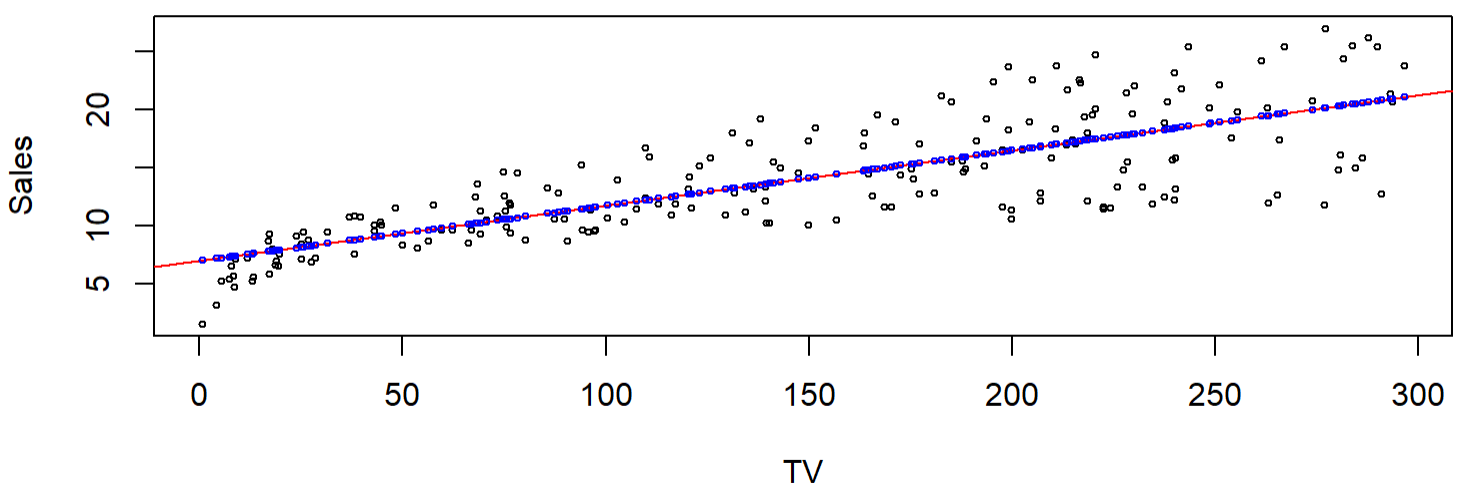
```
              Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Graphical representation of data (x_2, y) - black points, regression lines $(y = 7.033 + 0.047x_2)$ in red, and predicted values $(x, \hat{y})$ in blue;

```
1 with(Advertising, plot(TV, Sales, cex = 0.5))
2 abline(fitTV$coefficients, col = "red")
3 points(Advertising$TV, predict(fitTV), col = "blue", cex = 0.5)
```

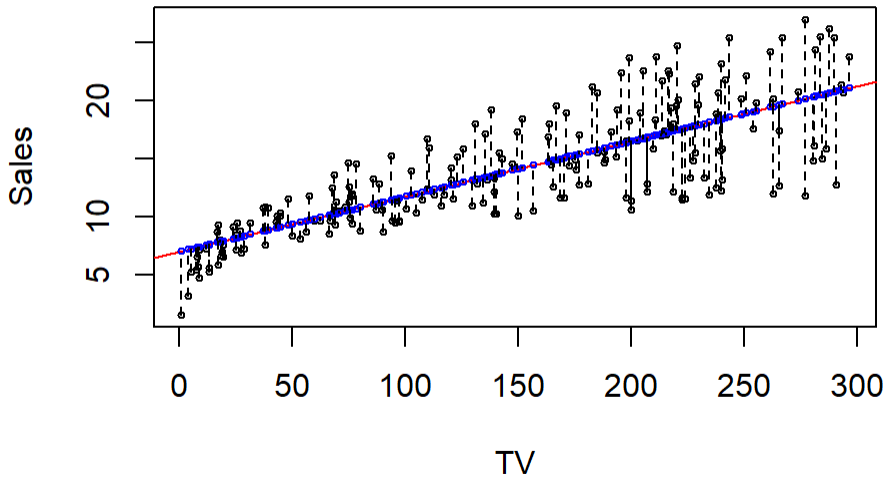


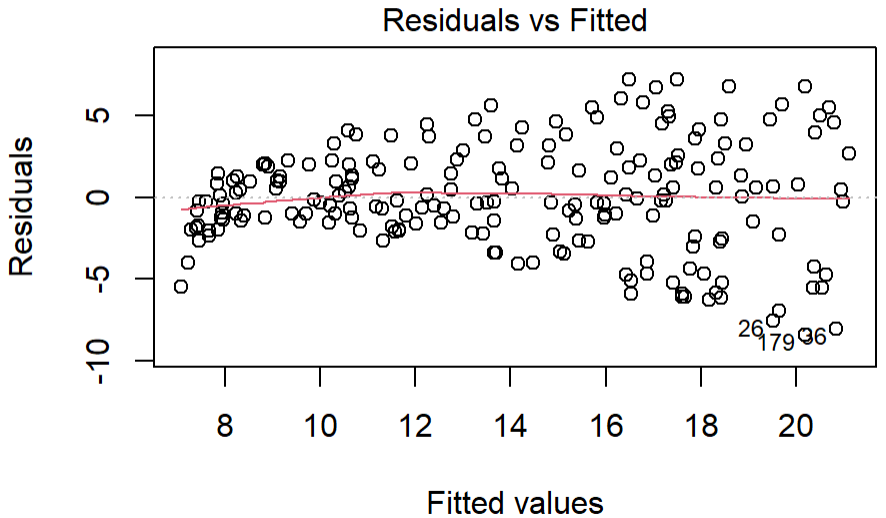
Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- The segments connecting y and $\hat{y}$ are composing the residuals $e = y - \hat{y}$

```
1 par(mfrow=c(1,2))
2 with(Advertising, plot(TV, Sales, cex = 0.5)); abline(fitTV$coefficients, col = "red
3 points(Advertising$TV, predict(fitTV), col = "blue", cex = 0.5)
4 with(Advertising, segments(x0 = TV, y0 = Sales, x1 = TV, y1 = predict(fitTV), lty =
5 plot(fitTV, which = 1)
```


Sales
5
10
20
0
50
100
150
200
250
300
TV


Residuals vs Fitted
Residuals
-10
-5
0
5
8
10
12
14
16
18
20
Fitted values
26
179
36

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Confidence interval:

$$IC_{\beta_r}^{1-\alpha} = (\hat{\beta}_r - t_{n-p;1-\alpha/2}\sqrt{S^2((X^\top X)^{-1}_{rr})}, \hat{\beta}_r + t_{n-p;1-\alpha/2}\sqrt{S^2((X^\top X)^{-1}_{rr})})$$

- For instance for β_1 $(1-\alpha = 0.95)$ a realization is

$$IC_{\beta_1}^{0.95} = (7.033 - t_{198;0.975} \times 0.458, 7.033 + t_{198;0.975} \times 0.458) \approx (6.13, 7.94)$$

- In absence of any TV advertising the sales, will, on average, fall somewhere between 6.13 and 7.94 units
- For each $1000\$$ increase in television advertising, there will be an average increase in sales betweem 42 and 53 units

```
1 confint(fitTV)
```

```
                2.5 %     97.5 %
(Intercept) 6.12971927 7.93546783
TV          0.04223072 0.05284256
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Hypothesis test to assess if there is a relationship between TV and Sales

$$\begin{cases} H_0: \text{There is no relationship between } x_2 \text{ and } y \\ H_1: \text{There is some relationship between } x_2 \text{ and } y \end{cases} \Longrightarrow \begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

$$t = \frac{\hat{\beta}_2}{\sqrt{S^2 (X^\top X)^{-1}_{22}}} \overset{H_0}{\sim} t_{n-p} \qquad p - value = \Pr(|t_{n-p}| > |t_r^{obs}|) = 2P(t_{n-p} \leq -|t_r^{obs}|)$$

- Here $t^{obs} = 0.04753664/0.002690607 = 17.66763$ and $\text{p} - \text{value} = 2\Pr(t_{198} \leq -17.66763)$
- We can infer that there is an association between TV advertising and sales

```
1  summary(fitTV)$coefficients[2,]
```

```
    Estimate   Std. Error      t value     Pr(>|t|)
4.753664e-02 2.690607e-03 1.766763e+01 1.467390e-42
```

```
1  2 * pt(-17.66763, df = 198)
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Assessing the accuracy of the model via residual standard error $s = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n} e_i^2}$
- Actual sales in each market deviate from the true regression line by approximately 3.26 units, on average (whether or not 3.26 units is acceptable or not depends on the problem context)
- Further, it is an absolute measure of the lack of fit (because it is measured in the units of y)
- If the predictions obtained using the model are very close to the true outcome values, that is if $\hat{y}_i \approx y_i$ for $i = 1, \ldots, n$, then s will be small and we can conclude that the model fits the data very well

```
1  summary(fitTV)$sigma
```

```
[1] 3.258656
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Assessing the accuracy of the model via R^2 coefficient (proportion of variability explained by the model)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- It does not depend on the scale of y: a value close to 1 means that a large proportion of the variability is explained by the regression, while a value close to 0 means that the regression line does not explain much of the variability of y (this might occur because the linear model is wrong and/or the error variance is high)
- Here, $R^2 = 0.61$ means the less than 2/3 of the variability in sales is explained by regressing sales on TV
- However, also in this case it is still challenging to determine what is a good R^2 value and it depends on the application (in typical applications in biology, psychology, marketing … the linear model is at best an extremely rough approximation to the data, and residual errors due to other unmeasured factors are often very large)

```
1  summary(fitTV)$r.squared
```

```
[1] 0.6118751
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

1. Considering the simple linear regression of sales on radio: A 1000 increasing in spending on radio advertising is associated with an increase of sales of around 203 units
2. Considering the simple linear regression of sales on newspaper: A 1000 increasing in spending on newspaper advertising is associated with an increase of sales of around 55 units

```r
fitRadio <- lm(Sales ~ Radio, data = Advertising)
summary(fitRadio)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
Radio       0.2024958 0.02041131  9.920765 4.354966e-19
```

```r
fitNewspaper <- lm(Sales ~ Newspaper, data = Advertising)
summary(fitNewspaper)$coefficients
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
Newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- However, we have not yet answered to the first question
- Indeed, the approach of fitting separate simple linear regression models for each predictor is not entirely satisfactory:

1. It is unclear how to make a single prediction of sales given the three advertising media budgets (each of the budgets is associated with a separate regression equation)
2. Each of three regression equations ignores the other two media in forming the estimates for the regression coefficients (since the media budgets are correlated with each other this can lead to very misleading estimates of the associaton between media budgets and sales)

```
1 round(cor(Advertising[,-1]),3)
```

```
             TV Radio Newspaper Sales
TV        1.000 0.055     0.057 0.782
Radio     0.055 1.000     0.354 0.576
Newspaper 0.057 0.354     1.000 0.228
Sales     0.782 0.576     0.228 1.000
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- In order to answer to the first question we must consider a multiple linear regression model where we regress sales onto TV, Radio, and Newspaper budgets, that is

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, for $i = 1, \ldots, n$, indepedently
- β_j for $j = 2, 3, 4$ quantifies the association between the predictor x_j and the response
- β_j is interpreted as the average effect of Y on a unit increase in x_j, holding all other predictors fixed

```
1  fitLM <- lm(Sales ~ TV + Radio + Newspaper, data = Advertising)
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Do you notice something different with respect to the simple linear regression models fitted above?

```
round(summary(fitLM)$coefficients,4)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.9389     0.3119  9.4223   0.0000
TV             0.0458     0.0014 32.8086   0.0000
Radio          0.1885     0.0086 21.8935   0.0000
Newspaper     -0.0010     0.0059 -0.1767   0.8599
```

```
summary(fitLM)$sigma
```

```
[1] 1.68551
```

```
summary(fitLM)$r.squared
```

```
[1] 0.8972106
```

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- For a given amount of

1. Radio and Newspaper advertising, spending an additional $1000\$$ on TV advertising is associated with additional 46 units of sales
2. TV and Newspaper advertising, spending an additional $1000\$$ on Radio advertising is associated with additional 189 units of sales
3. TV and Radio advertising, spending an additional $1000\$$ on Newspaper advertising is associated with 1 unit less of sales

- **While the regression coefficients for TV and Radio are almost the same of the simple linear regression models (0.0475 for TV and 0.2025 for Radio), the one for Newspaper (which was 0.0547 and significant) is now negative and no more significant (close to zero and p-value of 0.86)**

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- This because in the simple linear regression model, the slope term represents the average increase in sales associated with an additional $1000\$$ in Newspaper, ignoring TV and Radio, while in the multiple linear regression the slope term represents the average increase in sales associated with an additional $1000\$$ in Newspaper, holding fixed Radio and TV
- This is due to the correlation between Radio and Newspaper (0.35): markets with high Newspaper advertising tend to have high Radio advertising
- Indeed, in markets where we spend more on radio our sales will tend to be higher, and as the correlation shows, we also tend to spend more on newspaper advertising in those same markets
- In other words, Newspaper is a surrogate for Radio Advertising: newspaper gets *credit* for the association between radio and sales

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- Remember that relationship should not be interpreted as Cause-and-Effect; A statistical relationship — even a strong one — between and does not imply a cause-and-effect relationship
- Consider the example introduced in the first lecture: running a regression of shark attacks versus ice cream sales for data collected at a given beach in a community over a period of time would show a positive relationship
- Of course no one has (yet) suggested that ice creams should be banned at beaches to reduce shark attacks
- In reality, higher temperature causes more people to visit the beach, which in turn results in more ice cream sales and more shark attacks
- A multiple regression of shark attacks onto ice cream sales and temperatures reveals that ice cream sales is no longer significant after adjusting for the temperature

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- The question is equivalent to say:

1. Is there a relationship between the response and the predictors?
2. Is at least one of the predictors (x_2, x_3, x_4) useful in predicting the response?

- Basically we want test:

$$\begin{cases} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1 : \text{ at least one } \beta_j \text{ is different from zero} \end{cases}$$

- The hypothesis test is based on the F-statistic

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/(p-1)}{\hat{\sigma}^2/(n-p)} \overset{H_0}{\sim} F_{p-1,\,n-p}$$

where $\hat{\sigma}^2 = (\sum_{i=1}^{n} e_i^2)/n$ and $\tilde{\sigma}^2 = (\sum_{i=1}^{n} (y_i - \bar{y})^2)/n$ are used to get the observed test statistic

Linear Model: Example

1. Is there a relationship between advertising budget and sales?

- We have seen that running **summary(fitLM)**, the last raw reports the value of the observed F-statistic, along with the number of degrees of freedom for the numerator and the denominator, and the associated p-value
- Same information can be carried out using the **anova()** function
- When there is no relationship between y and the x's the value of the F-statistic is close to 1
- **However, instead to analyse the value of F, we can inspect the p-value: here the p-value is essentially zero, so we have extremely strong evidence that at least one predictor is associated with the response**

```
1 anova(lm(Sales ~ 1, data = Advertising), fitLM)
```

```
Analysis of Variance Table

Model 1: Sales ~ 1
Model 2: Sales ~ TV + Radio + Newspaper
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    199 5417.1
2    196  556.8  3    4860.3 570.27 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Model: Example

2. How strong is the relationship between advertising budget and sales?

- This question is equivalent to say: **How well does the model fit the data?** We can answer by using

1. The R^2 coefficient (% of variability explained by the model; square of the correlation between Y and $\hat{Y}$)

- The model using all the media as predictors get an R^2 of 0.8972, while the model including only TV and Radio has R^2 of 0.89719 (adding newspaper lead to a tiny increase of R^2: first signal to not consider newspaper in the model); instead we can see a large increase in R^2 when considering both TV and Radio than considering them separately; **the predictors explain around the 90% variance of the outcome**

```
summary(fitTV)$r.squared
```

```
[1] 0.6118751
```

```
summary(fitRadio)$r.squared
```

```
[1] 0.3320325
```

```
summary(lm(Sales ~ TV + Radio, data = Advertising))$r.squared
```

```
[1] 0.8971943
```

```
summary(fitLM)$r.squared # Model with all the variables
```

```
[1] 0.8972106
```

Linear Model: Example

2. How strong is the relationship between advertising budget and sales?

2. The Residual Standard Error $s = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n} e_i^2}$

- As before, we consider the model only including TV, the one including Radio, the one including TV and Radio, and the one including all the media: **the model including only TV and Radio seems to be more accurate (lower residual standard error)**; other evidence to not include Newspaper in the model
- The value is 1.69: each market deviate from the regression surface of about 1.69 units on average (note that the mean value is of 14.0 units)

```
summary(fitTV)$sigma
```

```
[1] 3.258656
```

```
summary(fitRadio)$sigma
```

```
[1] 4.274944
```

```
summary(lm(Sales ~ TV + Radio, data = Advertising))$sigma
```

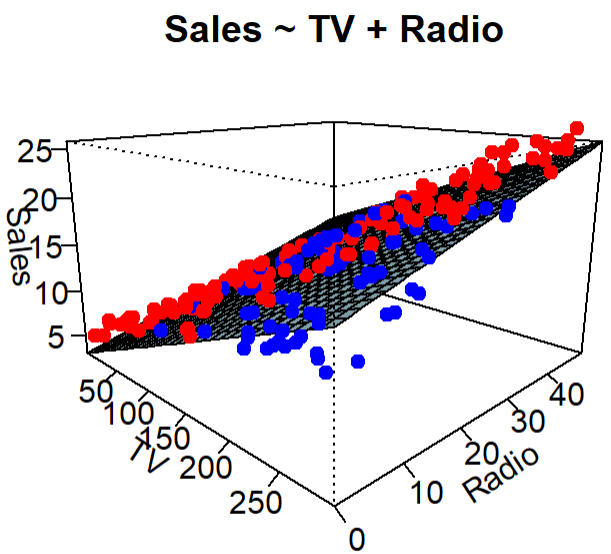
```
[1] 1.681361
```

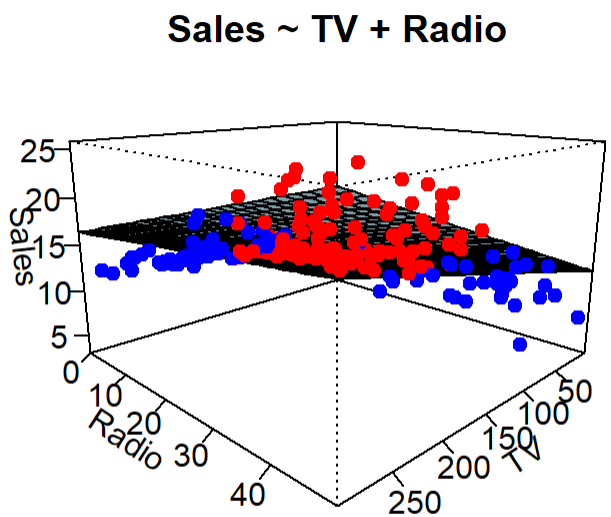
```
summary(fitLM)$sigma # Model with all the variables
```

Linear Model: Example

2. How strong is the relationship between advertising budget and sales?

- In addition to the R^2 coefficient and the residual standard error, it is interesting to analyze the model output graphically (sometimes numerical summaries are not enough to detect some aspects)
- Let's look to the 3d plot including the data and the regression surface from the fitted models: as expected some observations lie below and some above the regression plane; however, the linear model seems to overestimates Sales for instances in which most of the advertising money was spent exclusively in TV or Radio, while it underestimates for instances where the budget was split between the two media: this is probably due to an interaction effect (in red the points where $y > \hat{y}$ and in blue the points where $y < \hat{y}$)


Sales ~ TV + Radio
Sales
25
20
15
10
5
50
100
150
200
250
TV
0
10
20
30
40
Radio


Sales ~ TV + Radio
Sales
25
20
15
10
5
0
10
20
30
40
Radio
50
100
150
200
250
TV

Linear Model: Example

3. Which media are associated with sales?

- This question is equivalent to say: **Are all the three media associated with Sales, or are just one or two associated with Sales ?** To answer to this question, we must find a way to separate out the individual contribution of each medium to sales when we have spent money on all the three media?
  - After recognizing in point 1. that at least one is associated with the response, we must find the guilty one/ones
- **We could look at the the individual p-values: they are suggesting that the effect of Newspaper is no longer significant (alternative: perform variable selection; and pay attention when the number of covariates is large: the problem of false discoveries); this suggests that only TV and Radio are associated with Sales**

```
1  round(summary(fitLM)$coefficients,3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.939      0.312   9.422     0.00
TV              0.046      0.001  32.809     0.00
Radio           0.189      0.009  21.893     0.00
Newspaper      -0.001      0.006  -0.177     0.86
```

Linear Model: Example

4. How large is the association between each medium and sales?

- The standard error of $\hat{\beta}_j$ can be used to construct confidence intervals for β_j. The 95% confidence intervals for the coefficients are suggesting that: the confidence interval for TV and Radio are narrow and far from zero, providing evidence that these media are related to sales, while the intervals for newspaper includes zero, indicating that this variable is not statistically significant given the values of TV and Radio (there could be a collinearity problem?).

```
1 confint(fitLM)
```

```
                2.5 %     97.5 %
(Intercept)  2.32376228 3.55401646
TV           0.04301371 0.04851558
Radio        0.17154745 0.20551259
Newspaper   -0.01261595 0.01054097
```

Linear Model: Example

4. How large is the association between each medium and sales?

- In order to assess the association of each medium individually on the sales, we can use the three fitted simple linear regression models: strong association between TV and sales, as well between Radio and Sales, while there is a wild association between Newspaper and Sales, when TV and Radio are ignored

```
1 summary(fitTV)$coefficients
```

```
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

```
1 summary(fitRadio)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
Radio       0.2024958 0.02041131  9.920765 4.354966e-19
```

```
1 summary(fitNewspaper)$coefficients
```

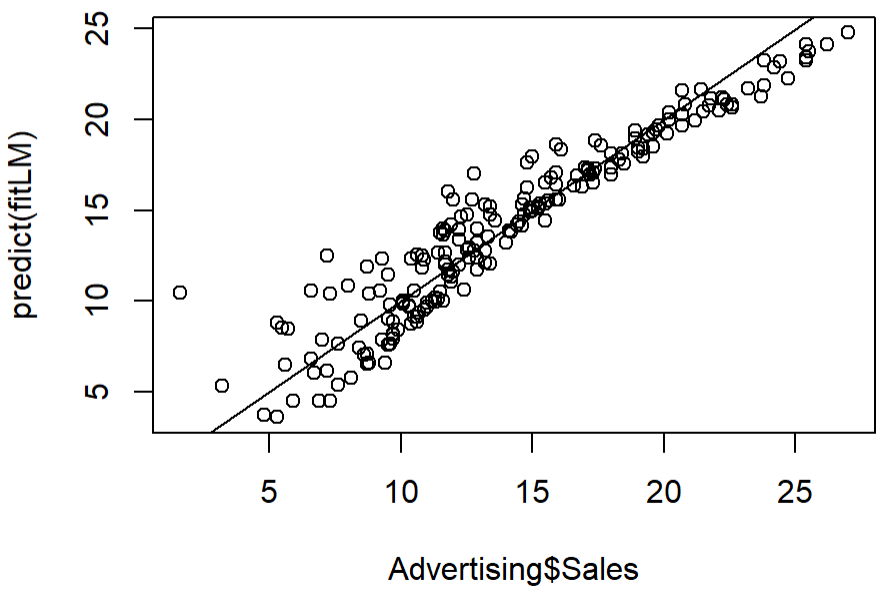
```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
Newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

Linear Model: Example

5. How well can we predict future sales?

- For any given level of television, radio and newspaper advertising, what is our prediction of sales, and what is the accuracy of this prediction?
- We can explore the accuracy by comparing y and $\hat{y}$: we can see from the plot that we are not doing a very good job

```
plot(Advertising$Sales, predict(fitLM))
abline(0,1)
```


predict(fitLM)
5
10
15
20
25
5
10
15
20
25
Advertising$Sales

Linear Model: Example

5. How well can we predict future sales?

- For any given level of television, radio and newspaper advertising, what is our prediction of sales, and what is the accuracy of this prediction?
- We can compute a confidence interval to quantify the uncertainty surrounding the average Sales, over a large number of cities
- For instance, given $10000\$$ spent on TV and $20000\$$ on Radio, the confidence interval is $[10.86, 11.71]$
- The 95% of the intervals of this form will contain the true population regression plane ($\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$)

```
predict(fitLM, se.fit = TRUE, interval = "confidence",
        newdata = data.frame(TV = 100, Radio = 20, Newspaper = 0))$fit
```

```
       fit      lwr      upr
1 11.28595 10.85908 11.71283
```

Linear Model: Example

5. How well can we predict future sales?

- We can compute a prediction interval to quantify the uncertainty surrounding the Sales for a particular city
- For instance, given $10000\$$ spent on TV and $20000\$$ on Radio, the prediction interval is $[7.93, 14.64]$
- The 95% of the intervals of this form will contain the true value of Y for this city (the interval is wider because we are accounting for the irreducibile error, that is the random error in the model)

```
predict(fitLM, se.fit = TRUE, interval = "predict",
        newdata = data.frame(TV = 100, Radio = 20, Newspaper = 0))$fit
```

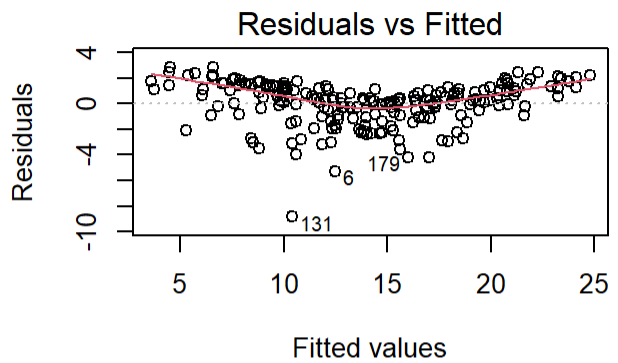
```
       fit      lwr      upr
1 11.28595 7.934592 14.63732
```

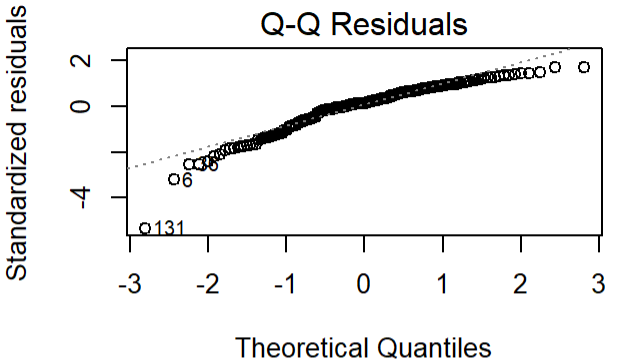
Linear Model: Example

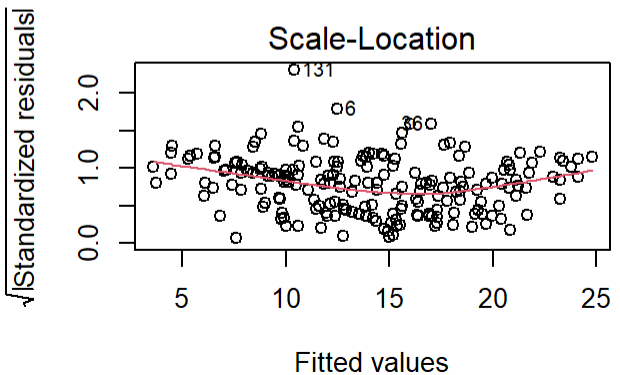
6. Is the relationship linear?

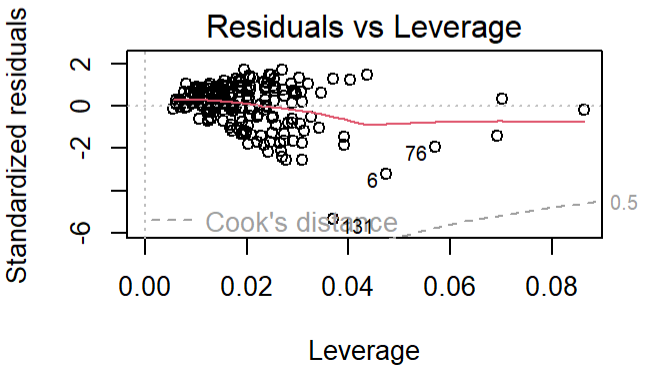
- The residuals plot can be used to identify non-linearity? If the relationships are linear the residuals plot(residuals vs fitted) should display no pattern
- The model shows several problems: residuals not uniformly distributed, substantial departure from normality, slight problem with the homoschedasticity assumption

```
1 par(mfrow=c(2,2))
2 plot(fitLM)
```


Residuals vs Fitted
Residuals
Fitted values
131
179
6


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
131
6


Scale-Location
√|Standardized residuals|
Fitted values
131
6


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
76
6
131

Linear Model: Example

6. Is the relationship linear?

- We have several possibility to address the problems detected by analyzing the resiudals:

1. Considering polynomial regression
2. Trasforming the outcome and/or the predictors

- Let's try to see what happens by considering quadratic terms of TV and Radio

```r
fitLMpoly <- lm(Sales ~ TV + Radio + I(TV^2), data = Advertising)
summary(fitLMpoly)
```

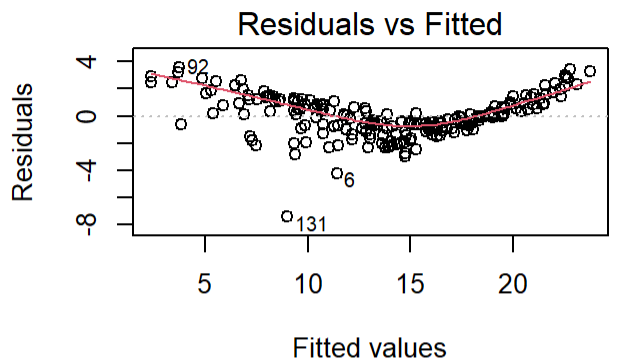
```
Call:
lm(formula = Sales ~ TV + Radio + I(TV^2), data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3860 -0.8822 -0.0498  0.9613  3.5725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.288e+00  3.588e-01   3.588 0.000421 ***
```

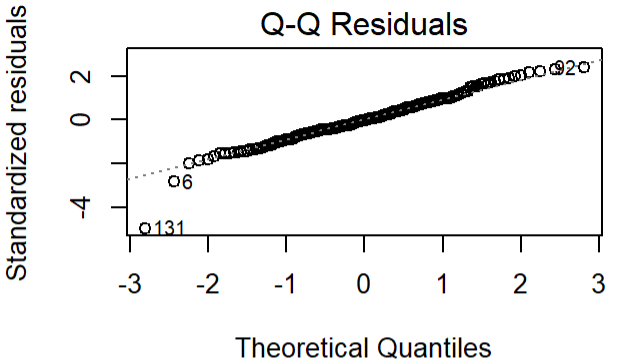
```
TV           7.844e-02  4.985e-03  15.736  < 2e-16 ***
Radio        1.930e-01  7.293e-03  26.465  < 2e-16 ***
I(TV^2)     -1.136e-04  1.677e-05  -6.775 1.42e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

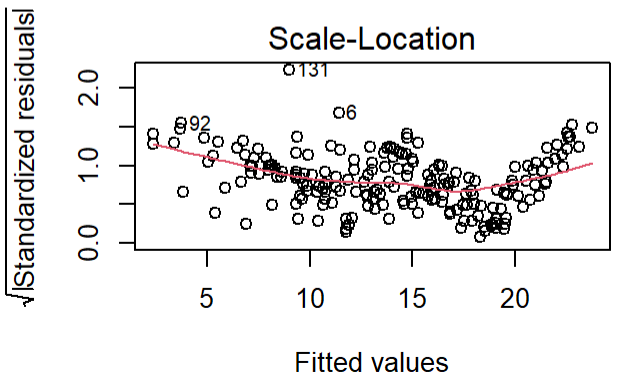
Linear Model: Example

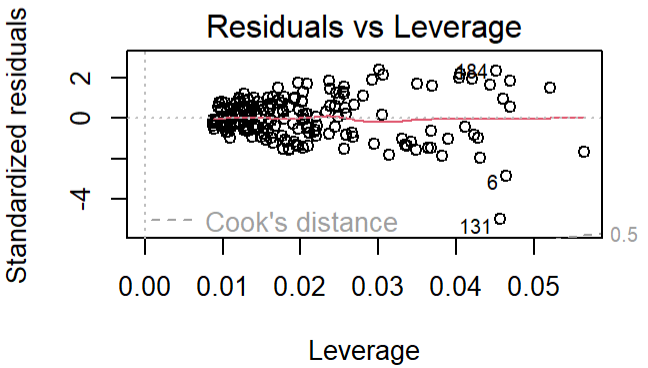
> **6. Is the relationship linear?**
>
> - It seems that the non-linearity patter in the residuals cannot be solved by simply considering the quadratic term (we should address differently)

```
1  par(mfrow=c(2,2))
2  plot(fitLMpoly)
```


Residuals vs Fitted
Residuals
Fitted values
92
6
131


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
92
6
131


Scale-Location
√|Standardized residuals|
Fitted values
131
6
92


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
484
6
131

Linear Model: Example

7. Is there any synergy among the advertising media?

- We could ask whether spending $50000\$$ on TV and $50000\$$ on Radio is associated to higher sales than spending $100000\$$ to either television or radio individually. In marketing, it is known as *sinergy* effect, while in statistics we refer to *interaction* effect
- We concluded that both TV and Radio seems to be associated to Sales
- The linear model

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

  assumes that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media
- The 3d plots suggested that an interaction effect may be present (we noticed that when TV and Radio are low, then the true sales are lower than the values predicted by the linear model)

Linear Model: Example

7. Is there any synergy among the advertising media?

- Then, we fit the model including the interaction effect $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2} \times x_{i3} + \varepsilon_i$
- We can interpret β_4 as

1. The effectiveness of TV advertising with a one-unit increase in radio advertising
2. The effectiveness of Radio advertising with a one-unit increase in TV advertising Indeed

- Adding $1000\$$ on Radio will increase the number of Sales of $\beta_2 + \beta_4 \times TV$
- Adding $1000\$$ on TV will increase the number of Sales of $\beta_3 + \beta_4 \times Radio$

Linear Model: Example

7. Is there any synergy among the advertising media?

- The results obtained by fitting the model suggest that it is sensible to include the interaction term (p-value extremely low indicating that there is evidence for $H_1 : \beta_4 \neq 0$)
- The R^2 coefficient is 0.968, showing a marked increase with respect to the model not including the interaction term (without interaction 0.897)
- An increase of $1000\$$ on TV is associated with an increase in the number of Sales of $19 + 1.1 \times Radio$ units
- An increase of $1000\$$ on Radio is associated with an increase in the number of Sales of $29 + 1.1 \times TV$ units

```
fitLMint <- lm(Sales ~ TV*Radio, data = Advertising)
# Equivalently
fitLMint <- lm(Sales ~ TV + Radio + TV:Radio, data = Advertising)
round(summary(fitLMint)$coefficients,4)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.7502     0.2479 27.2328   0.0000
TV             0.0191     0.0015 12.6990   0.0000
Radio          0.0289     0.0089  3.2408   0.0014
TV:Radio       0.0011     0.0001 20.7266   0.0000
```

```
summary(fitLMint)$r.squared
```

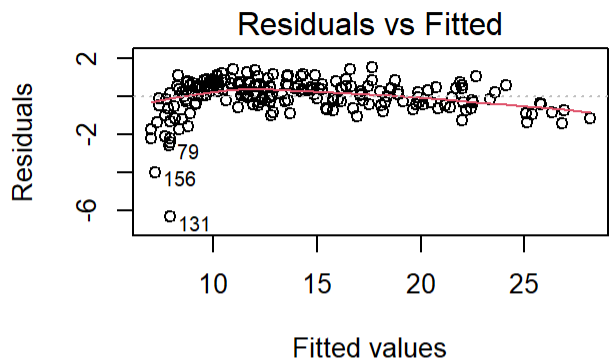
```
[1] 0.9677905
```

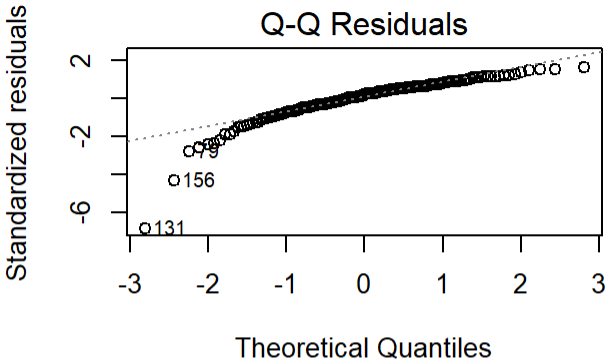
Linear Model: Example

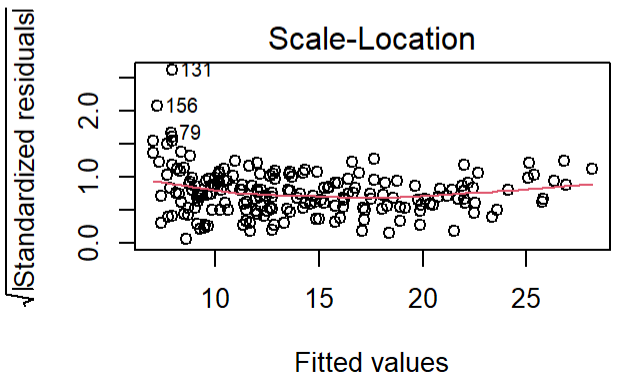
7. Is there any synergy among the advertising media?

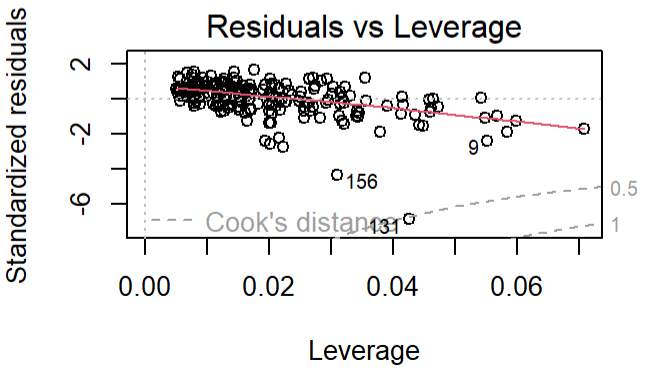
- However, the Residuals plot shows a little improvement, although there is something that we are missing

```
par(mfrow=c(2,2))
plot(fitLMint)
```


Residuals vs Fitted
Residuals
Fitted values
79
156
131


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
79
156
131


Scale-Location
√|Standardized residuals|
Fitted values
131
156
79


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
1
156
9
131

Linear Model

Transforming the outcome

- Let's try to see what happens if we transform the outcome variable
- Note that this R^2 cannot be compared with ones obtained without trasforming the outcome (we should report the values on the original scale and computing it)

```
1 fitLMtrans <- lm(I(log(Sales)) ~ TV + I(TV^2) +  Radio + TV:Radio, data = Advertisir
2 summary(fitLMtrans)
```

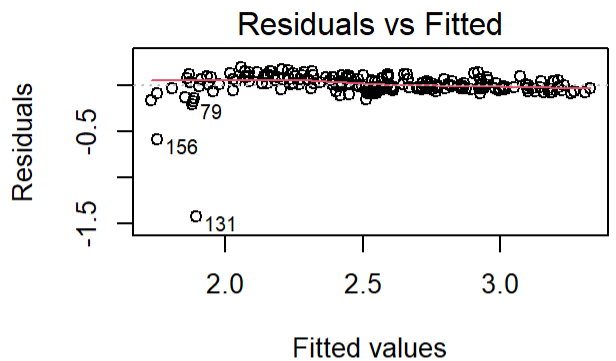
```
Call:
lm(formula = I(log(Sales)) ~ TV + I(TV^2) + Radio + TV:Radio,
    data = Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42334 -0.03854 -0.00468  0.06355  0.19313

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.651e+00  4.111e-02  40.157  < 2e-16 ***
TV           7.703e-03  4.763e-04  16.174  < 2e-16 ***
I(TV^2)     -1.798e-05  1.471e-06 -12.223  < 2e-16 ***
Radio        5.960e-03  1.259e-03   4.735 4.21e-06 ***
```

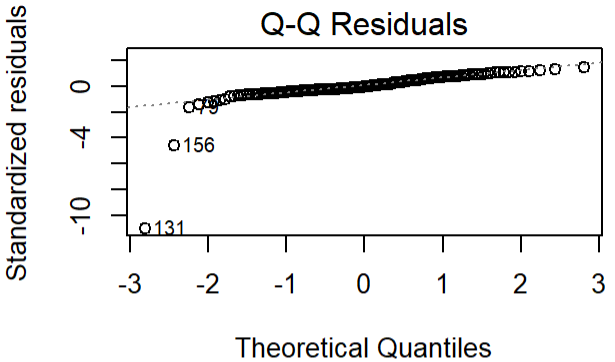
Linear Model

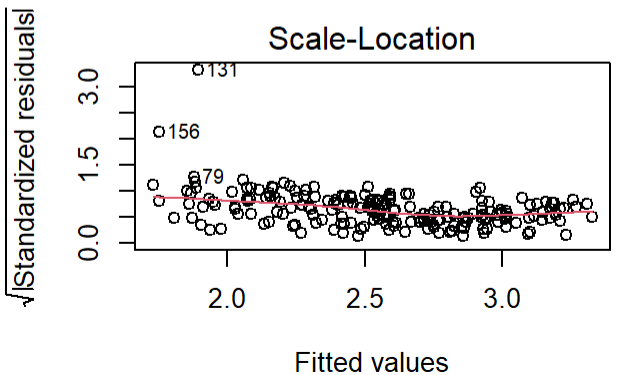
Transforming the outcome

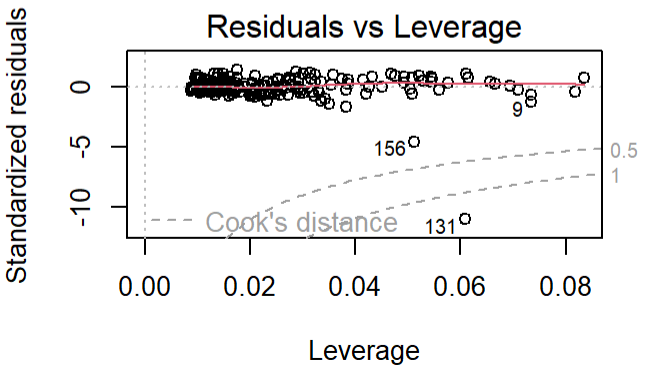
- The Residuals plot are suggesting that considering the logarithm of the outcome could be a viable option (but also considering transformation of the predictors)

```
1 par(mfrow=c(2,2))
2 plot(fitLMtrans)
```


Residuals vs Fitted
Residuals
Fitted values
79
156
131


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
156
131


Scale-Location
√|Standardized residuals|
Fitted values
131
156
79

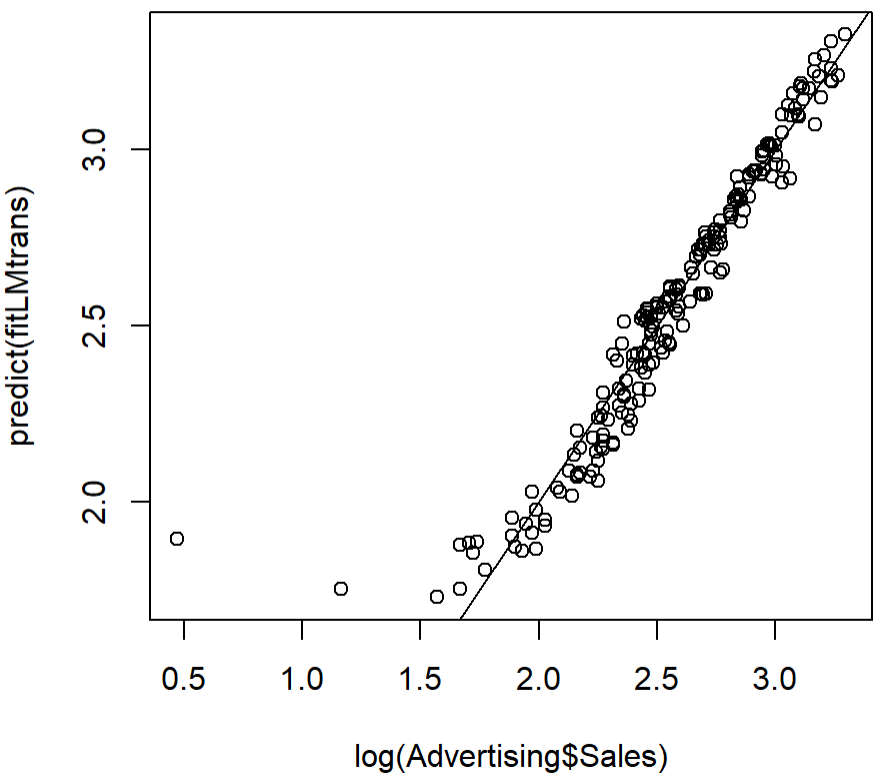

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
1
9
156
131

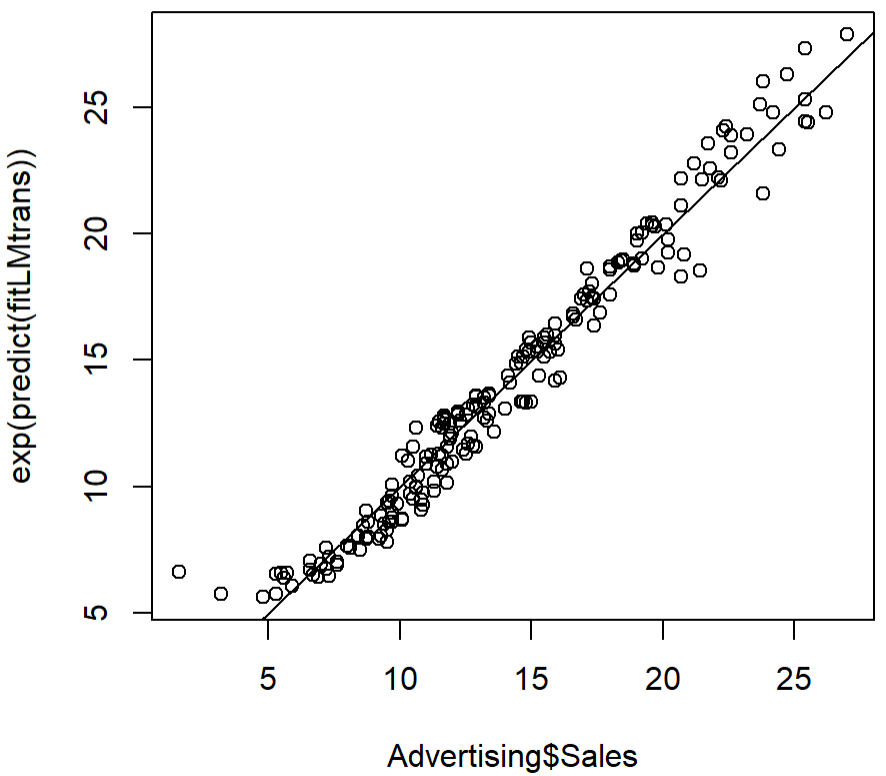
Linear Model

Transforming the outcome

- Predicted values in the two scales

```
par(mfrow = c(1,2))
plot(log(Advertising$Sales), predict(fitLMtrans))
abline(0,1)
plot(Advertising$Sales, exp(predict(fitLMtrans)))
abline(0,1)
```


predict(fitLMtrans)
2.0
2.5
3.0
0.5
1.0
1.5
2.0
2.5
3.0
log(Advertising$Sales)


exp(predict(fitLMtrans))
5
10
15
20
25
5
10
15
20
25
Advertising$Sales

Linear Model

What we will explore next week?

- Non constant variance of the error terms
- Correlation of Error Terms
- Multicollinearity problem
- Endogeneity
- Instrumental variable estimator
- Some tools for causal inference