Laborator

Elemente de regresie liniară simplă

Obiectivul acestui laborator este de a prezenta câteva exemple legate de problema de regresie liniară simplă.

1 Introducere

Regresia liniară simplă (sau *modelul liniar simplu*) este un instrument statistic utilizat pentru a descrie relația dintre două variabile aleatoare, x (variabilă *cauză*, *predictor* sau *covariabilă*) și y (variabilă *răspuns* sau *efect*) și este definit prin

$$\mathbb{E}[y|x] = \beta_0 + \beta_1 x$$

sau altfel spus

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

În relațiile de mai sus, β_0 și β_1 sunt cunoscute ca ordonata la origine (*intercept*) și respectiv panta (*slope*) dreptei de regresie.

Ipotezele modelului sunt:

i. **Linearitatea**: $\mathbb{E}[y|x] = \beta_0 + \beta_1 x$

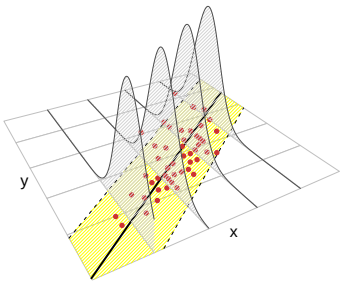
ii. **Homoscedasticitatea**: $Var(\varepsilon_i) = \sigma^2$, cu σ^2 constantă pentru $i = 1, \ldots, n$

iii. **Normalitatea**: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ pentru $i = 1, \ldots, n$

iv. **Independența erorilor**: $\varepsilon_1, \ldots, \varepsilon_n$ sunt independente (sau necorelate, $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$, $i \neq j$, deoarece sunt presupuse normale)

Altfel spus

$$y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$



- Nicio ipoteză nu a fost făcută asupra repartiției lui X (poate fi sau deterministă sau aleatoare)
- Modelul de regresie presupune că Y **este continuă** datorită normalității erorilor. În orice caz, X **poate fi o variabilă discretă**!

Dat fiind un eșantion $(x_1, y_1), \ldots, (x_n, y_n)$ pentru variabilele z și y putem estima coeficienții necunoscuți β_0 și β_1 minimizând *suma abaterilor pătratice reziduale* (*Residual Sum of Squares* - RSS)

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

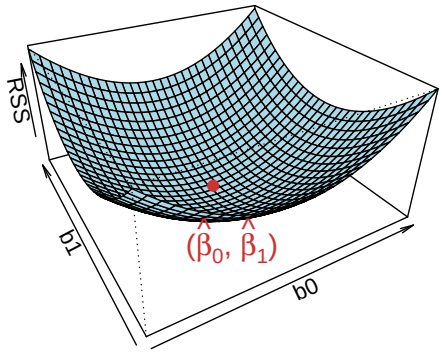
ceea ce conduce la

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

unde folosim notațiile

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ este *media eșantionului*
- $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ este *suma abaterilor pătratice pentru* x
- $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ este *suma abaterilor pătratice pentru* y
- $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ este *suma produselor încrucișate*

Graficul functiei RSS pentru modelul $y = -0.5 + 1.5x + e$:



Odată ce avem estimatorii $(\hat{\beta}_0, \hat{\beta}_1)$, putem defini:

- *valorile ajustate* (*fitted values*) $\hat{y}_1, \ldots, \hat{y}_n$ (valorile verticale pe dreapta de regresie), unde

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

- *valori reziduale* (*estimated residuals*) $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ (distanțele verticale dintre punctele actuale (X_i, Y_i) și cele ajustate la model $(X_i, \hat{Y}_i)$), unde

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

Estimatorul pentru σ^2 este

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}.$$

1.1 Funcția *lm()* din R

Pentru a rula modelul de regresie liniară simplă în R se folosește funcția `lm()` (*linear model*). Funcția `lm()` are două argumente esențiale: `formula` și `data`.

Argument	Descriere
`formula`	O formulă de forma $y \sim x_1 + x_2 + \ldots$, unde y este variabila răspuns (dependentă) iar $x_1, x_2, \ldots$ sunt variabilele explicative (covariabilele). Dacă vrem să includem toate coloanele (cu excepția lui y) ca variabile explicative putem folosi $y \sim .$
`data`	Este setul de date în format `data.frame` care conține coloanele specificate de formulă.

Următorul tabel conține corespondențe între codul `R` și conceptele statistice asociate modelului de regresie:

`R`	Concepte statistice
`x`	Variabilele predictor $x_1, \ldots, x_n$
`y`	Răspunsul $y_1, \ldots, y_n$
`data <- data.frame(x = x, y = y)`	Eșantionul $(x_1, y_1), \ldots, (x_n, y_n)$
`model <- lm(y ~ x, data = data)`	Modelul de regresie liniară simplă
`model$coefficients`	Coeficienții $\hat{\beta}_0, \hat{\beta}_1$
`model$residuals`	Valorile reziduale $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$
`model$fitted.values`	Valorile ajustate (fitate) $\hat{y}_1, \ldots, \hat{y}_n$
`model$df.residual`	Gradele de libertate $n-2$
`summaryModel <- summary(model)`	Sumarul modelului de regresie liniară
`summaryModel$sigma`	Estimatorul $\hat{\sigma}$
`summaryModel$r.squared`	Coeficientul de determinare R^2
`summaryModel$fstatistic`	Testul lui Fisher F
`anova(model)`	Tabelul ANOVA

2 Aplicație

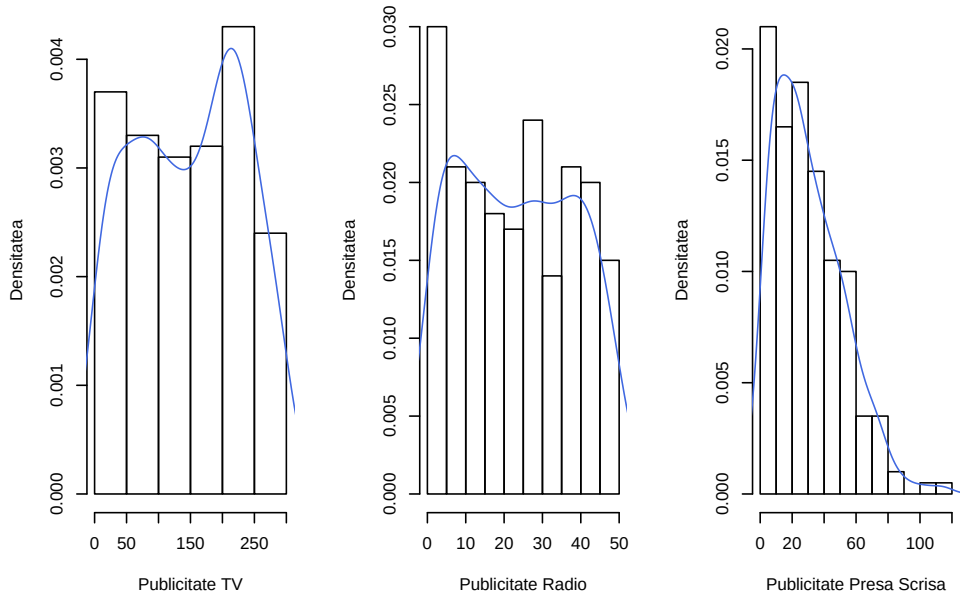
Ne propunem să investigăm relația dintre volumul vânzărilor dintr-un anumit produs (calculate în mii de unități) și bugetul (în mii de RON) alocat pentru publicitatea la acestuia prin trei canale media: televizor, radio și presa scrisă. Pentru aceasta vom folosi setul de date advertising care conține informații despre volumul vânzărilor și bugetul alocat publicității în 200 de piețe de desfacere.

Începem prin a înregistra și sumariza setul de date

```
advertise = read.csv("dataIn/advertising.csv", row.names = 1)

summary(advertise)
       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  
```

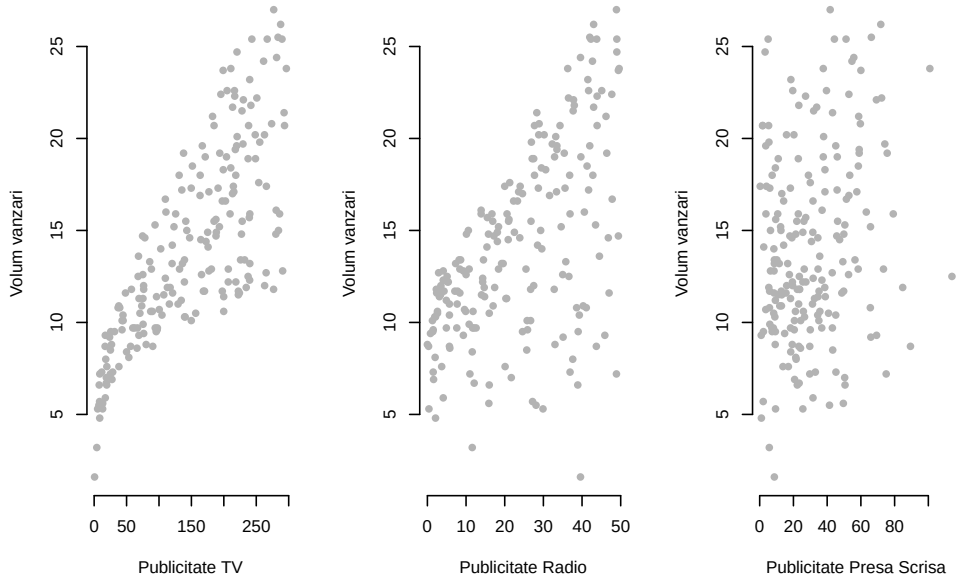
unde observăm că volumul de vânzări (`sales`) variază între 1.6 și 27 de mii de unități pe când bugetul alocat publicității TV variază între 0.7 și 296.4 mii de RON cu o medie de 147.04 mii RON. Putem vedea cum sunt distribuite bugetele alocate publicității trasând histogramele acesto variabile:

```
par(mfrow = c(1,3))

hist(advertise$TV,
     probability = TRUE,
     main = "",
     cex.main = 0.7,
     xlab = "Publicitate TV",
     ylab = "Densitatea")
lines(density(advertise$TV),
      col = "royalblue")

hist(advertise$radio,
     probability = TRUE,
     main = "",
     cex.main = 0.7,
     xlab = "Publicitate Radio",
     ylab = "Densitatea")
lines(density(advertise$radio),
      col = "royalblue")

hist(advertise$newspaper,
     probability = TRUE,
     main = "",
     cex.main = 0.7,
     xlab = "Publicitate Presa Scrisa",
     ylab = "Densitatea")
lines(density(advertise$newspaper),
      col = "royalblue")
```



Pentru a ilustra relația dintre volumul de vânzări în funcție de bugetul alocat fiecărui canal de publicitate trasăm diagramele de împrăștiere corespunzătoare:

```
par(mfrow = c(1,3))

plot(advertise$TV, advertise$sales,
     xlab = "Publicitate TV",
     ylab = "Volum vanzari",
     col = "grey70",
     pch = 16,
     bty="n")

plot(advertise$radio, advertise$sales,
     xlab = "Publicitate Radio",
     ylab = "Volum vanzari",
     col = "grey70",
     pch = 16,
     bty="n")

plot(advertise$newspaper, advertise$sales,
     xlab = "Publicitate Presa Scrisa",
     ylab = "Volum vanzari",
     col = "grey70",
     pch = 16,
     bty="n")
```

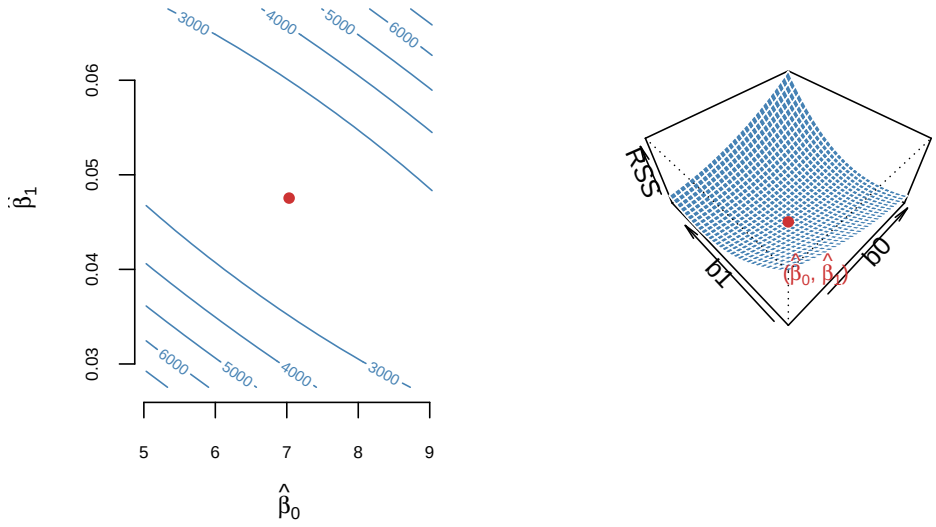


În analiza următoare ne vom opri doar asupra relației dintre volumul de vânzări și bugetul alocat publicității la TV.

2.1 Estimarea parametrilor

Considerăm modelul de regresie liniară simplă $y = \beta_0 + \beta_1 x + \varepsilon$ (unde x =`advertise$TV` iar y =`advertise$sales`), $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, a cărui parametrii sunt β_0, β_1 și σ^2. Asta înseamnă că modelăm volumul de vânzări mediu ca o funcție liniară de bugetul alocat publicității TV.

Din graficul sumei abaterilor pătratice reziduale RSS (pentru graficul din stânga am folosit funcția `contour` iar pentru cel din dreapta funcția `persp`)



observăm că funcția este convexă și admite un punct de minim.

Estimatorii parametrilor β_0 și β_1 obținuți prin metoda celor mai mici pătrate sunt dați de

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{și} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

care se traduc în `R` prin

```
# pentru beta1

b1 = cov(advertise$TV, advertise$sales)/var(advertise$TV)
cat("b1 = ", b1, "\n")
b1 =  0.04753664

# sau

sum((advertise$TV-mean(advertise$TV))*(advertise$sales))/
  sum((advertise$TV-mean(advertise$TV))^2)
[1] 0.04753664

# pentru beta0

b0 = mean(advertise$sales) - b1*mean(advertise$TV)
cat("b0 = ", b0)
b0 =  7.032594
```

sau folosind funcția `lm()`:

```
# inregistram modelul (o lista)
advertise_TV_model = lm(sales~TV, data = advertise)

# afisam elementele listei
```

```
names(advertise_TV_model)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

Pentru a ilustra valorile coeficienților folosim

```
advertise_TV_model$coefficients
(Intercept)          TV
 7.03259355  0.04753664
```

Observăm că modelul obținut este

$$\text{Volum vânzări}_i = 7.033 + 0.047 \times \text{ Buget Publicitate TV}_i + \varepsilon_i$$

ceea ce arată că dacă nu se alocă niciun buget publicității atunci volmul mediu de vânzări este de 7033 de unități iar dacă bugetul alocat publicității TV crește cu 1000 de RON atunci volumul de vânzări crește în medie cu 47 de unități.

Dacă ne interesăm la matricea de varianță-covarianță a acestora obținem

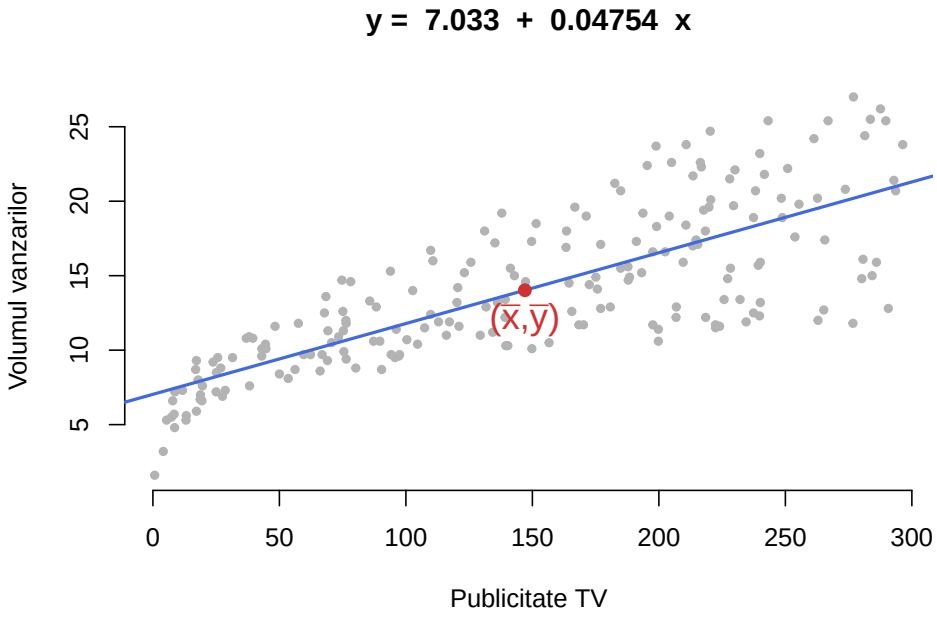
$$W = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} & -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}$$

iar pentru a o determina numeric folosim funcția `vcov()`:

```
W = vcov(advertise_TV_model)
W
             (Intercept)            TV
(Intercept)  0.209620158 -1.064495e-03
TV          -0.001064495  7.239367e-06
```

În figura de mai jos vom ilustra că dreapta de regresie trece prin punctul coordonate $(\bar{x}, \bar{y})$:

```
plot(advertise$TV, advertise$sales,
     xlab = "Publicitate TV",
     ylab = "Volumul vanzarilor",
     col = "grey70",
     pch = 20,
     bty="n",
     main = paste("y = ", format(b0, digits = 4), " + ",
                  format(b1, digits = 4), " x"))

abline(a = b0, b = b1, col = "royalblue", lwd = 2)
points(mean(advertise$TV), mean(advertise$sales),
       pch = 16,
       col = "brown3",
       cex = 1.2)
text(mean(advertise$TV), mean(advertise$sales)-2,
     col = "brown3", cex = 1.4,
     labels = expression(paste("(", bar(x), ",", bar(y),")")))
```



De asemenea pentru calculul estimatorului lui σ ($\hat{\sigma}$) avem

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

și în R

```
n = length(advertise$sales)
e_hat = advertise$sales - (b0+b1*advertise$TV)

rss = sum(e_hat^2)
# sau folosind comanda deviance(advertise_TV_model)

sigma_hat = sqrt(rss/(n-2))
sigma_hat
[1] 3.258656
```

sau cu ajutorul funcției `lm()`

```
sqrt(deviance(advertise_TV_model)/df.residual(advertise_TV_model))
[1] 3.258656
```

sau încă aplicând funcția `summary()`

```
advertise_TV_model_summary = summary(advertise_TV_model)

advertise_TV_model_summary$sigma
[1] 3.258656
```

2.2 Intervale de încredere pentru parametrii

Repartițiile lui $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right), \quad \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

unde

$$\sigma^2_{\hat{\beta}_0} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right], \quad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}.$$

Trebuie menționat că în practică $\sigma_{\hat{\beta}_0}$ și $\sigma_{\hat{\beta}_1}$ se mai notează și prin $\text{SE}(\hat{\beta}_0)^2$ și respectiv $\text{SE}(\hat{\beta}_1)^2$ (SE - standard error).

Folosind estimatorul $\hat{\sigma}^2$ pentru σ^2 obținem că

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2}, \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$

unde

$$\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right], \quad \hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{S_{xx}}$$

prin urmare, intervalele de încredere de nivel $1-\alpha$ pentru β_0 și β_1 sunt

$$IC = \left(\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} t_{n-2;1-\alpha/2}\right), \quad j = 0, 1.$$

În R putem calcula aceste intervale de încredere folosind comenzile

```
alpha = 0.05

# trebuie avut grija ca functia var si sd se calculeaza
# impartind la (n-1) si nu la n !!!

se_b0 = sqrt(sigma_hat^2*(1/n+mean(advertise$TV)^2/((n-1)*var(advertise$TV))))
se_b1 = sqrt(sigma_hat^2/((n-1)*var(advertise$TV)))

lw_b0 = b0 - qt(1-alpha/2, n-2)*se_b0
up_b0 = b0 + qt(1-alpha/2, n-2)*se_b0

cat("CI pentru b0 este (", lw_b0, ", ", up_b0, ")\n")
CI pentru b0 este ( 6.129719 ,  7.935468 )

lw_b1 = b1 - qt(1-alpha/2, n-2)*se_b1
up_b1 = b1 + qt(1-alpha/2, n-2)*se_b1

cat("CI pentru b1 este (", lw_b1, ", ", up_b1, ")")
CI pentru b1 este ( 0.04223072 ,  0.05284256 )
```

Același rezultat se obține apelând funcția `confint()` :

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

Observăm că în absența unui buget de publicitate TV, volumul vânzărilor se încadrează în medie cu o încredere de 95% între 6130 și 7940 de unități. Mai mult cu același nivel de încredere, pentru fiecare creștere a bugetului pentru publicitatea TV cu 1000 de RON obținem o creștere, în medie, a vânzărilor între 42 și 53 de unități.

Putem construi și o regiune de încredere de nivel de încredere $1-\alpha$ pentru vectorul $\boldsymbol{\beta} = (\beta_0, \beta_1)$ plecând de la repartiția bidimensională a vectorului $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^\intercal$. În cazul modelului condițional normal avem că $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 V)$ unde

$$V = \frac{1}{\sigma^2}\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})}\begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

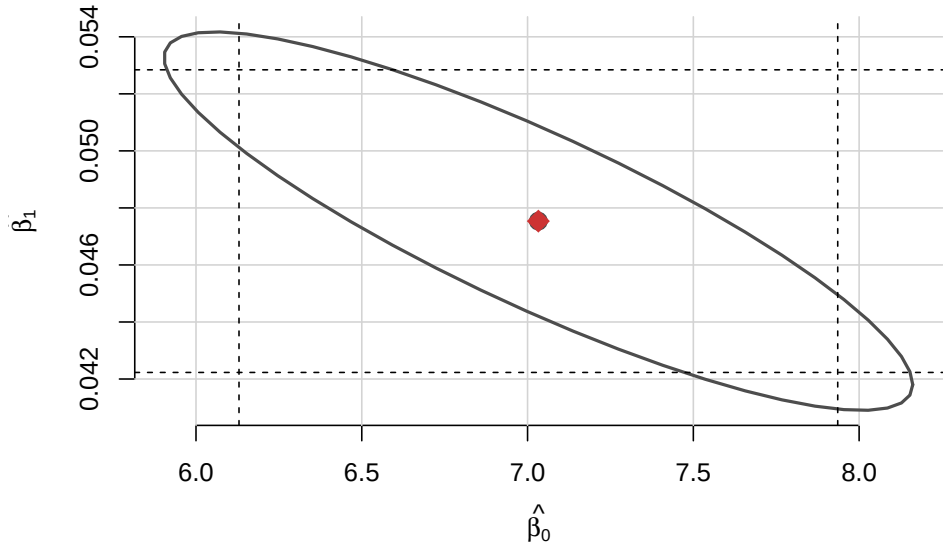
și cum

$$\frac{1}{2\hat{\sigma}^2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\intercal V^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim F_{2,n-2}$$

găsim că regiunea de încredere este

$$RC(\beta_0, \beta_1) = \left\{ \frac{1}{2\hat{\sigma}^2}\left[n(\hat{\beta}_0 - \beta_0)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n x_i^2 \right] \leq f_{2,n-2}^{1-\alpha} \right\}$$

```
library(ellipse)
library(car)

par(bty = "n")

# trasam regiunea de incredere
confidenceEllipse(advertise_TV_model,
                  xlab = expression(hat(beta[0])),
                  ylab = expression(hat(beta[1])),
                  col = "grey30")

points(coef(advertise_TV_model)[1], coef(advertise_TV_model)[2],
       pch = 18, col = "brown3",
       cex = 2)

# trasam intervalele de incredere
abline(v = confint(advertise_TV_model)[1,], lty = 2)
abline(h = confint(advertise_TV_model)[2,], lty = 2)
```



aceasta evidențiind corelația negativă dintre parametrii.

2.3 ANOVA pentru regresie

În această secțiune ne propunem să răspundem la întrebarea: Este predictorul x folositor în prezicerea răspunsului y ? cu alte cuvinte vrem să testăm ipoteza nulă $H_0: \beta_1 = 0$.

Introducem următoarele *sume de abateri pătratice*:

- $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$, **suma totală a abaterilor pătratice** (variația totală a lui $y_1, \ldots, y_n$).
- $SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, **suma abaterilor pătratice de regresie** (variabilitatea explicată de dreapta de regresie)
- $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, **suma abaterilor pătratice reziduale**

Avem următoarea descompunere ANOVA

$$\underbrace{SS_T}_{\text{Variația lui } Y_i} = \underbrace{SS_{reg}}_{\text{Variația lui } \hat{Y}_i} + \underbrace{RSS}_{\text{Variația lui } \hat{\varepsilon}_i}$$

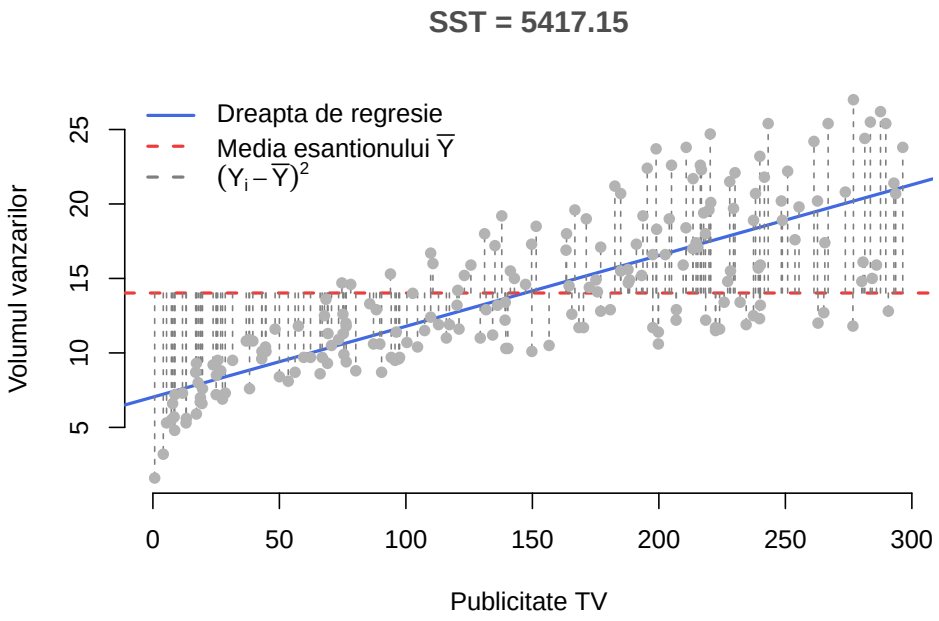
și tabelul ANOVA corespunzător

	Df	SS	MS	F	p-value
Predictor	1	SS_{reg}	$\frac{SS_{reg}}{1}$	$\frac{SS_{reg}/1}{RSS/(n-2)}$	p
Residuuri	$n-2$	RSS	$\frac{RSS}{n-2}$		

Descompunerea ANOVA pentru problema noastră poate fi ilustrată astfel:

a) *suma abaterilor pătratice totală*:

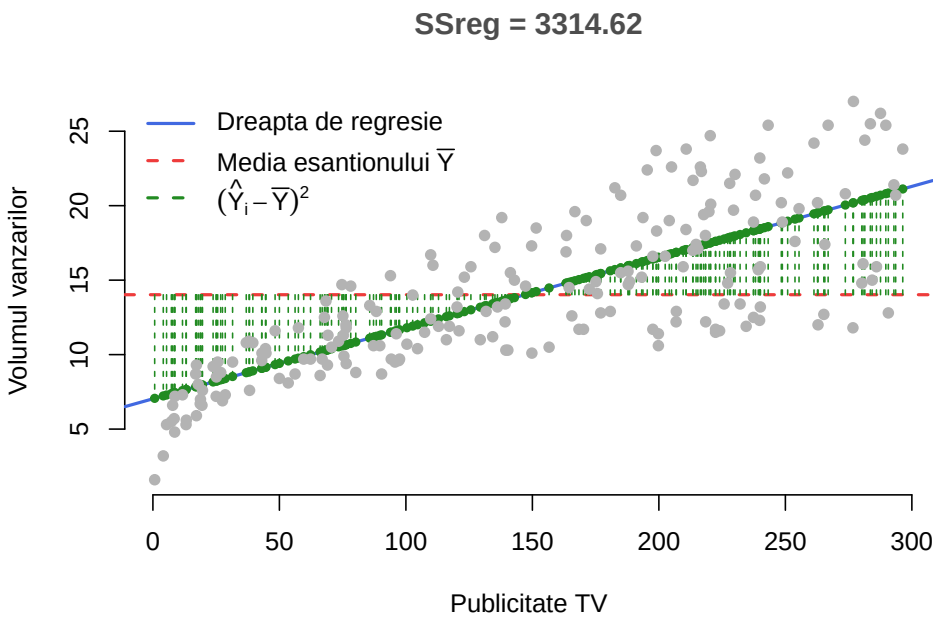
```
plot(advertise$TV, advertise$sales, pch = 20, type = "n",
     main = paste("SST =", round(sum((advertise$sales -
                                        mean(advertise$sales))^2), 2)),
     col.main = "grey30",
     xlab = "Publicitate TV",
     ylab = "Volumul vanzarilor",
     bty = "n")

abline(advertise_TV_model$coefficients, col = "royalblue", lwd = 2)
abline(h = mean(advertise$sales), col = "brown2", lty = 2, lwd = 2)

segments(x0 = advertise$TV, y0 = mean(advertise$sales),
         x1 = advertise$TV, y1 = advertise$sales,
         col = "grey50", lwd = 1, lty = 2)

legend("topleft",
       legend = expression("Dreapta de regresie", "Media esantionului " * bar(Y),
                           (Y[i] - bar(Y))^2),
       lwd = c(2, 2, 2),
       col = c("royalblue", "brown2", "grey50"),
       lty = c(1, 2, 2),
       bty = "n")

points(advertise$TV, advertise$sales, pch = 16, col = "grey70")
```



b) *suma abaterilor pătratice de regresie*

```
plot(advertise$TV, advertise$sales, pch = 20, type = "n",
     main = paste("SSreg =",
                  round(sum((advertise_TV_model$fitted.values -
                                mean(advertise$sales))^2), 2)),
     col.main = "grey30",
```

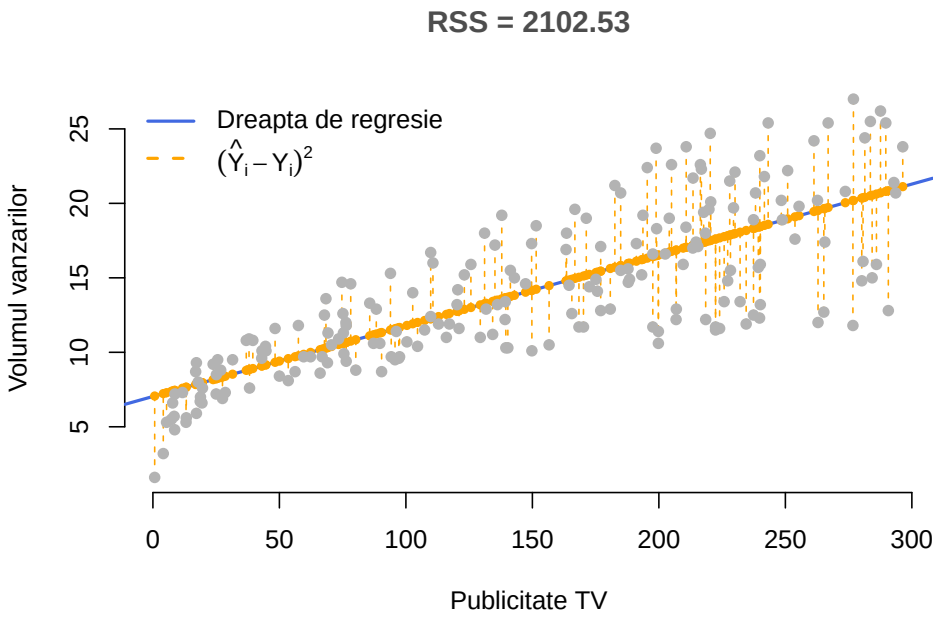
```
     xlab = "Publicitate TV",
     ylab = "Volumul vanzarilor",
     bty = "n")

abline(advertise_TV_model$coefficients, col = "royalblue", lwd = 2)
abline(h = mean(advertise$sales), col = "brown2", lty = 2, lwd = 2)

segments(x0 = advertise$TV, y0 = mean(advertise$sales),
         x1 = advertise$TV, y1 = advertise_TV_model$fitted.values,
         col = "forestgreen", lwd = 1, lty = 2)

points(advertise$TV, advertise_TV_model$fitted.values, pch = 20, col = "forestgreen")

legend("topleft",
       legend = expression("Dreapta de regresie", "Media esantionului " * bar(Y),
                                        (hat(Y)[i] - bar(Y))^2),
       lwd = c(2, 2, 2),
       col = c("royalblue", "brown2", "forestgreen"),
       lty = c(1, 2, 2),
       bty = "n")

points(advertise$TV, advertise$sales, pch = 16, col = "grey70")
```



c) *suma abaterilor pătratice reziduale*

```
plot(advertise$TV, advertise$sales, pch = 20, type = "n",
     main = paste("RSS =",
                  round(sum((advertise$sales - advertise_TV_model$fitted.values)^2), 2)),
     col.main = "grey30",
     xlab = "Publicitate TV",
     ylab = "Volumul vanzarilor",
     bty = "n")
```

```
abline(advertise_TV_model$coefficients, col = "royalblue", lwd = 2)

segments(x0 = advertise$TV, y0 = advertise$sales,
         x1 = advertise$TV, y1 = advertise_TV_model$fitted.values,
         col = "orange", lwd = 1, lty = 2)

points(advertise$TV, advertise_TV_model$fitted.values, pch = 20, col = "orange")

legend("topleft",
       legend = expression("Dreapta de regresie", (hat(Y)[i] - Y[i])^2),
       lwd = c(2, 2),
       col = c("royalblue", "orange"),
       lty = c(1, 2),
       bty = "n")

points(advertise$TV, advertise$sales, pch = 16, col = "grey70")
```



Tabelul ANOVA se obține prin

```
# tabel ANOVA
anova(advertise_TV_model)
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)
TV          1 3314.6  3314.6  312.14 < 2.2e-16 ***
Residuals 198 2102.5    10.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Definiția *coeficientului de determinare* R^2 este strâns legată de descompunerea ANOVA:

$$R^2 = \frac{SS_{reg}}{SS_T} = \frac{SS_T - RSS}{SS_T} = 1 - \frac{RSS}{SS_T}$$

R^2 măsoară **proporția din variația** variabilei răspuns y **explicată** de variabila predictor x prin regresie. Proporția din variația totală a lui y care nu este explicată este $1 - R^2 = \frac{RSS}{SS_T}$. Intuitiv, R^2 măsoară cât de bine modelul de regresie este în concordanță cu datele (cât de strâns este norul de puncte în jurul dreptei de regresie). Observăm că dacă datele concordă *perfect* cu modelul (adică $RSS = 0$) atunci $R^2 = 1$.

În cazul problemei noastre avem $R^2 =$0.612 prin urmare aproximativ 61.19% din variabilitatea volumului vânzărilor este explicată de bugetul alocat publicității TV.

Putem vedea că $R^2 = r_{xy}^2$, unde r_{xy} este *coeficientul de corelație* empiric:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

```
cor(advertise$TV, advertise$sales)^2
[1] 0.6118751
```

Mai mult se poate verifica și că $R^2 = r_{y\hat{y}}^2$, adică *coeficientul de determinare este egal cu pătratul coeficientului de corelație empirică dintre* $y_1, \ldots, y_n$ *și* $\hat{y}_1, \ldots, \hat{y}_n$.

Verificăm relația $R^2 = r_{xy}^2 = r_{y\hat{y}}^2$ numeric:

```
yHat = advertise_TV_model$fitted.values

advertise_TV_model_summary$r.squared # R^2
[1] 0.6118751
cor(advertise$TV, advertise$sales)^2 # corelatia^2 dintre x si y
[1] 0.6118751
cor(advertise$sales, yHat)^2 # corelatia^2 dintre y si yHat
[1] 0.6118751
```

2.4 Inferență asupra parametrilor

Este predictorul X folositor în prezicerea răspunsului Y ? Vrem să testăm ipoteza nulă $H_0 : \beta_j = 0$ (pentru $j = 1$ spunem că predictorul `nivel de sare` nu are un efect *liniar* semnificativ asupra `tensiunii arteriale`). Pentru aceasta vom folosi statistica de test

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \underset{H_0}{\sim} t_{n-2}.$$

Funcția `summary` ne întoarce p-valoarea corespunzătoare a acestor teste:

```
summary(advertise_TV_model)

Call:
lm(formula = sales ~ TV, data = advertise)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Tabelul de sub cuvântul `coefficients` care apare afișat în urma rulării funcției `summary()` este modul tipic de afișare a rezultatelor rulării unui model de regresie (nu numai în `R`). Acesta prezintă pentru fiecare linie (coeficient) 5 coloane: prima (`Estimate`) conține valorile etimate ale coeficienților modelului de regresie (care apar ca denumire a liniilor); a doua (`Std. Error`) reprezintă abaterile standard ale coeficienților estimați ($\hat{\sigma}_{\hat{\beta}_j}$); a treia (`t value`) prezintă valoarea observată a statisticii de test ($\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$) a cărui ipoteză nulă este $H_0 : \beta_j = 0$ versus alternativa $H_1 : \beta_j \neq 0$; a patra coloană (`Pr(>|t|)`) reprezintă probabilitatea critică (*p-valoarea*), i.e. probabilitatea ca stastistica de test sub ipoteza nulă să depășească valoarea estimată. Ultima coloană afișează o versiune grafică a testului, de exemplu $***$ reprezintă că testul respinge ipoteza nulă H_0 pentru erori de primă speță mai mari sau egale cu 0.001. Dacă ipoteza nulă nu este respinsă atunci trebuie să reconsiderăm modelul. De asemenea, tabelul prezintă și valoarea estimatorului lui σ, aici $\hat{\sigma} = 3.259$, a gradelor de libertate $n - 2 = 198$ și a coeficientului de determinare $R^2 = 0.6119$ (`Multiple R-squared`)[1]. Ultima linie, folositoare în special în cazul modelului de regresie multiplă, indică valoarea statisticii de test (`F`) și respectiv p-valoarea asociată testului de comparare a modelului folosit și cel în care apare doar termenul β_0, celelalte fiind nule. În cazul modelului de regresie liniară simplă avem relația $F = t_1^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right)^2$.

Observăm că ambele ipoteze sunt respinse în favoarea alternativelor bilaterale (la aceeași concluzie am ajuns și utitându-ne la intervalele de încredere - nu conțineau valoarea 0). Putem observa că t_1^2 este exact valoarea F statisticii, deci cele două abordări ne dau aceleași rezultate numerice.

2.4.1 Predicție

Pentru un nou set de predictori, x_0, răspunsul prognozat este $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ și vrem să investigăm incertitudinea din această predicție. Putem face distincția între două tipuri de predicție: predicție asupra răspunsului viitor mediu (inferență asupra mediei condiționate $\mathbb{E}[y|x_0]$) sau predicție asupra observațiilor viitoare (inferență asupra răspunsului condiționat $y|x_0$).

Un interval de încredere pentru răspunsul viitor mediu este:

$$\left(\hat{y} \pm t_{n-2;1-\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)$$

Un interval de încredere pentru valoarea prezisă (interval de predicție) este:

$$\left(\hat{y} \pm t_{n-2;1-\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)$$

Pentru a găsi aceste intervale vom folosi funcția `predict()`:

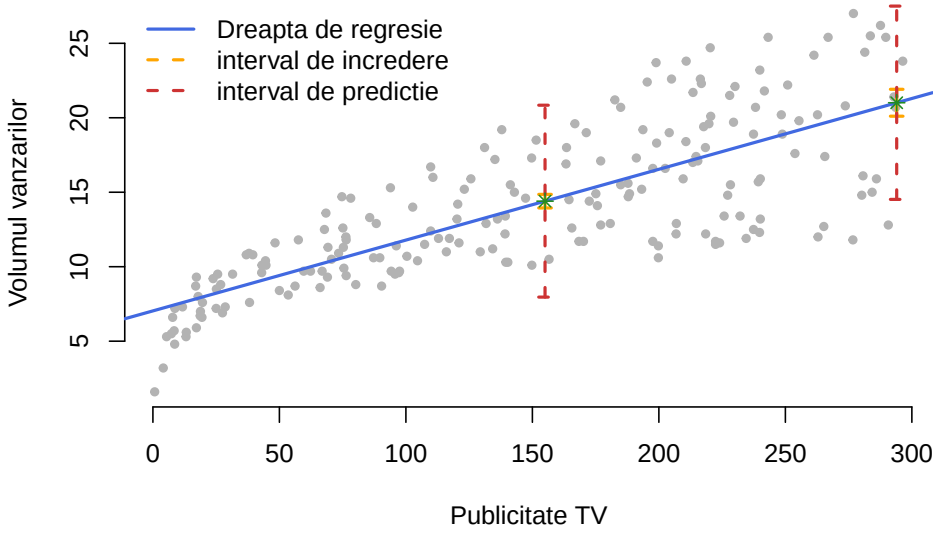
[1] Coeficientul de detertminare ajustat R_a este definit prin $R_a = 1 - \frac{n-1}{n-2}\frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbf{1}\|^2}$

```
newData = data.frame(TV = 150)
newData2 = data.frame(TV = c(130, 140, 150))

# Predictie
predict(advertise_TV_model, newdata = newData)
       1
14.16309

# Predictie pentru valoarea raspunsului mediu
predict(advertise_TV_model, newdata = newData, interval = "confidence")
       fit      lwr      upr
1 14.16309 13.70842 14.61776
predict(advertise_TV_model, newdata = newData2, interval = "confidence")
       fit      lwr      upr
1 13.21236 12.74905 13.67566
2 13.68772 13.23179 14.14365
3 14.16309 13.70842 14.61776

# Predictie asupra observatiilor viitoare
predict(advertise_TV_model, newdata = newData, interval = "prediction")
       fit      lwr      upr
1 14.16309 7.720898 20.60528
predict(advertise_TV_model, newdata = newData2, interval = "prediction")
       fit      lwr      upr
1 13.21236 6.769550 19.65516
2 13.68772 7.245442 20.13000
3 14.16309 7.720898 20.60528
```

Volumul de vânzări prezis pentru o anumită valoare x_0 impreună cu intervalul de încredere de nivel 95% pentru răspunsul mediu și cu intervalul de predicție, sunt ilustrate în figura următoare

```
alpha = 0.05
x0 = c(155, 294)

p.conf = predict(advertise_TV_model, data.frame(TV = x0), se = T, interval = "confidence")
p.pred = predict(advertise_TV_model, data.frame(TV = x0), se = T, interval = "prediction")

# diagrama de imprastiesre
plot(advertise$TV, advertise$sales,
     col = "grey70", pch = 20,
     xlab = "Publicitate TV",
     ylab = "Volumul vanzarilor",
     bty = "n")

# dreapta de regresie
abline(advertise_TV_model$coefficients, col = "royalblue", lwd = 2)

#intervalele de incredere
segments(x0 = x0, y0 = p.conf$fit[,2], x1 = x0, y1 = p.conf$fit[,3],
         col = "orange", lty = 1, lwd = 2)
segments(x0 = x0-2.5, y0 = p.conf$fit[,2], x1 = x0+2.5, y1 = p.conf$fit[,2],
         col = "orange", lty = 1, lwd = 2)
segments(x0 = x0-2.5, y0 = p.conf$fit[,3], x1 = x0+2.5, y1 = p.conf$fit[,3],
         col = "orange", lty = 1, lwd = 2)
```

```
#intervalele de predictie
segments(x0 = x0, y0 = p.pred$fit[,2], x1 = x0, y1 = p.pred$fit[,3],
         col = "brown3", lty = 2, lwd = 2)
segments(x0 = x0-2.5, y0 = p.pred$fit[,2], x1 = x0+2.5, y1 = p.pred$fit[,2],
         col = "brown3", lty = 2, lwd = 2)
segments(x0 = x0-2.5, y0 = p.pred$fit[,3], x1 = x0+2.5, y1 = p.pred$fit[,3],
         col = "brown3", lty = 2, lwd = 2)

# valoarea prezisa
points(x0, p.conf$fit[,1],
       col = "forestgreen",
       pch = 8)

legend("topleft",
       legend = c("Dreapta de regresie",
                  "interval de incredere",
                  "interval de predictie"),
       lwd = c(2, 2, 2),
       col = c("royalblue", "orange", "brown3"),
       lty = c(1, 2, 2),
       bty = "n")
```



Sunt circumstanțe în care am dori să avem intervalele de încredere pentru răspunsul mediu în mai mult de un punct, prin urmare ne aflăm în cadrul unei probleme de inferență simultană. O soluție la această problemă, în cazul în care avem m puncte, este dată de inegalitatea lui Bonferroni care conduce la marginea

$$\hat{\beta}_0 + \hat{\beta}_1 x_{0i} - t_{n-2,\frac{\alpha}{2m}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{0i} - \bar{x})^2}{S_{xx}}} < \beta_0 + \beta_1 x_{0i} < \hat{\beta}_0 + \hat{\beta}_1 x_{0i} + t_{n-2,\frac{\alpha}{2m}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{0i} - \bar{x})^2}{S_{xx}}}.$$

Scheffe (a se vedea [Casella and Berger, 2001, pag. 559 - 562]) a arătat, pentru problema de regresie, că

există un interval care este adevărat pentru orice x

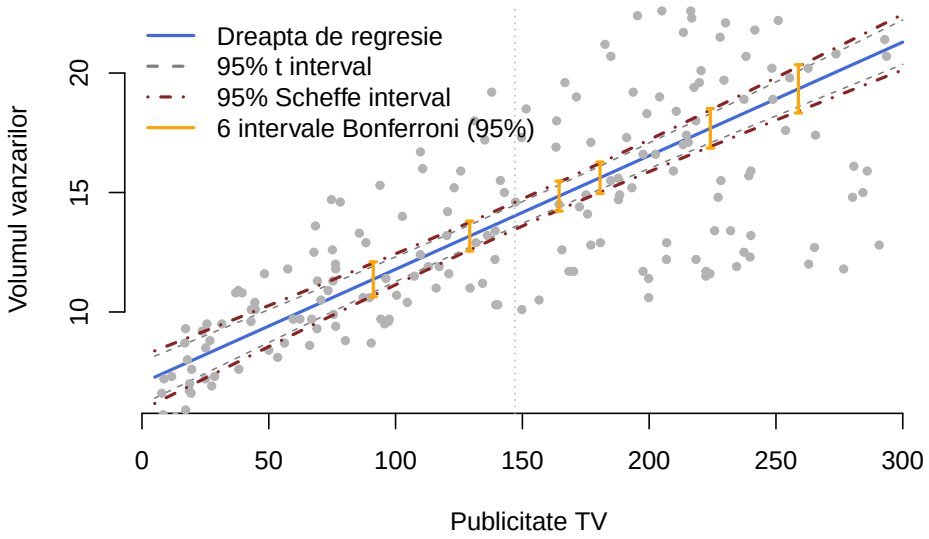
$$\hat{\beta}_0 + \hat{\beta}_1 x - M_\alpha \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} < \beta_0 + \beta_1 x < \hat{\beta}_0 + \hat{\beta}_1 x + M_\alpha \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}, \forall x$$

unde $M_\alpha = \sqrt{2f^\alpha_{2,n-2}}$.

În figura de mai jos am ilustrat volumul de vânzări prezis împreună cu intervalul de încredere de nivel 95% pentru răspunsul mediu folosind 6 intervale de tip Bonferroni precum și banda lui Scheffe:

```
alpha = 0.05
g = seq(5,300,0.5)

p = predict(advertise_TV_model, data.frame(TV = g), se = T, interval = "confidence")

matplot(g, p$fit, type = "l", lty = c(1,2,2),
        lwd = c(2,1,1),
        col = c("royalblue", "grey50", "grey50"),
        xlab = "Publicitate TV",
        ylab = "Volumul vanzarilor",
        bty = "n")

# rug(advertise$TV)

points(advertise$TV, advertise$sales,
       col = "grey70", pch = 20)
abline(v = mean(advertise$TV), lty = 3, col = "grey65")

# Scheffe's bounds
M = sqrt(2*qf(1-alpha, 2, n-2))

s_xx = (n-1)*var(advertise$TV)
lw_scheffe = b0 + b1*g - M*sigma_hat*sqrt(1/n+(g-mean(advertise$TV))^2/s_xx)
up_scheffe = b0 + b1*g + M*sigma_hat*sqrt(1/n+(g-mean(advertise$TV))^2/s_xx)

lines(g, lw_scheffe,
      lty = 4,
      lwd = 2,
      col = "brown4")
lines(g, up_scheffe,
      lty = 4,
      lwd = 2,
      col = "brown4")

# Bonferroni bounds

x0 = runif(6, min = 10, max = 290)
m = length(x0)

t_bonf = qt(1-alpha/(2*m), n-2)

lw_bonf = b0 + b1*x0 - t_bonf*sigma_hat*sqrt(1/n+(x0-mean(advertise$TV))^2/s_xx)
up_bonf = b0 + b1*x0 + t_bonf*sigma_hat*sqrt(1/n+(x0-mean(advertise$TV))^2/s_xx)
```

```
segments(x0 = x0, y0 = lw_bonf, x1 = x0, y1 = up_bonf,
         col = "orange", lty = 1, lwd = 2)
segments(x0 = x0-1.25, y0 = lw_bonf, x1 = x0+1.25, y1 = lw_bonf,
         col = "orange", lty = 1, lwd = 2)
segments(x0 = x0-1.25, y0 = up_bonf, x1 = x0+1.25, y1 = up_bonf,
         col = "orange", lty = 1, lwd = 2)

legend("topleft", legend = c("Dreapta de regresie", "95% t interval",
                                    "95% Scheffe interval",
                             paste0(m, " intervale Bonferroni (95%)")),
       lwd = c(2, 2, 2, 2),
       col = c("royalblue", "grey50", "brown4", "orange"),
       lty = c(1, 2, 4, 1),
       bty = "n")
```



2.5 Diagnostic

În această secțiune vom vedea dacă setul nostru de date verifică ipotezele modelului de regresie liniară.

a) *Independența*

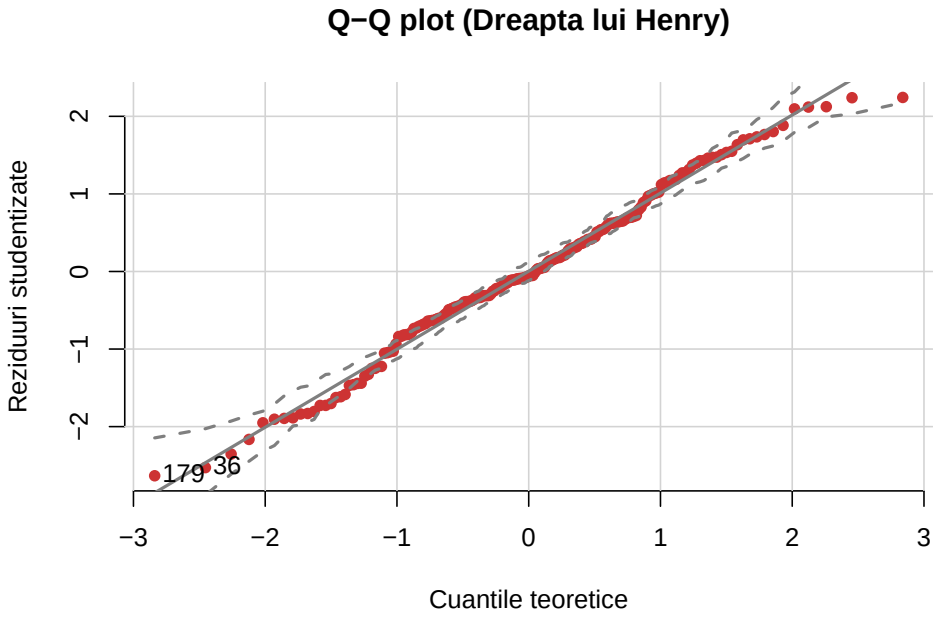
Ipoteza de independență a variabilei răspuns (prin urmare și a erorilor) reiese, de cele mai multe ori, din modalitatea în care s-a desfășurat experimentul.

b) *Normalitatea*

Pentru a verifica dacă ipoteza de normalitate a erorilor este satisfăcută vom trasa dreapta lui Henry (sau Q-Q plot-ul):

```
# library(car)
par(bty = "n")
qqPlot(advertise_TV_model, col = "brown3", col.lines = "grey50", pch = 16,
```

```
       simulate = TRUE,
       xlab = "Cuantile teoretice",
       ylab = "Reziduuri studentizate",
       main = "Q-Q plot (Dreapta lui Henry)",
       bty = "n")
```




```
[1]  36 179
```

Putem folosi și testul `Shapiro-Wilk`:

```
shapiro.test(residuals(advertise_TV_model))

    Shapiro-Wilk normality test

data:  residuals(advertise_TV_model)
W = 0.99053, p-value = 0.2133
```

c) *Homoscedasticitatea*

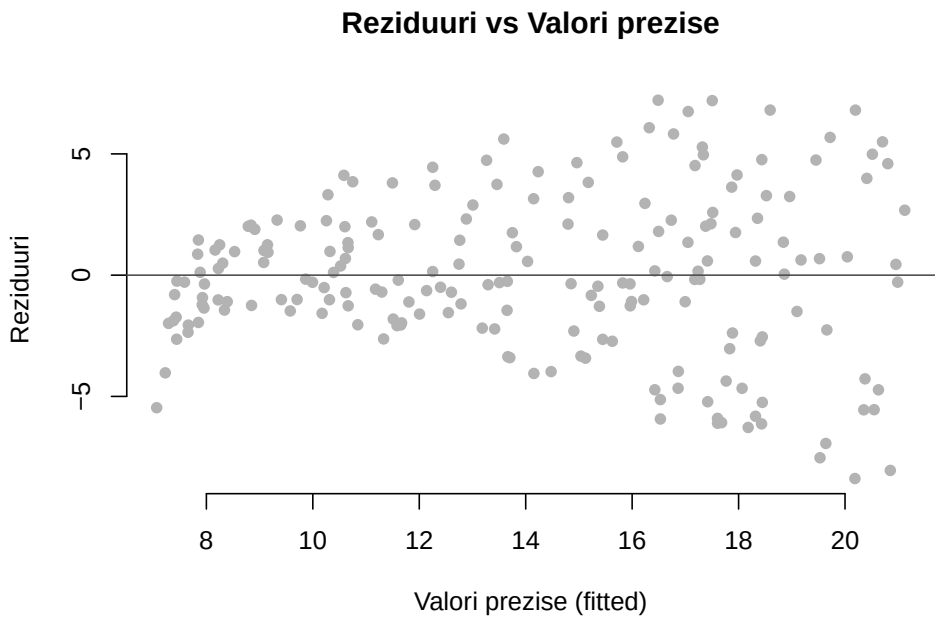
Pentru a verifica proprietatea de homoscedasticitate a erorilor vom trasa un grafic al reziduurilor versus valorile prezise (fitted), i.e. $\hat{\varepsilon}$ vs $\hat{y}$. Dacă avem homoscedasticitate a erorilor atunci ar trebui să vedem o variație constantă pe verticală ($\hat{\varepsilon}$).

Tot în acest grafic putem observa dacă ipoteza de liniaritate este verificată (în caz de liniaritate între variabila răspuns și variabila explicativă nu are trebui să vedem o relație sistematică între reziduuri și valorile prezise - ceea ce nu se întâmplă în cazul nostru) ori dacă există o altă legătură structurală între variabila dependentă (răspuns) și cea independentă (predictor).

În cazul aplicației noastre avem următoarea figură:

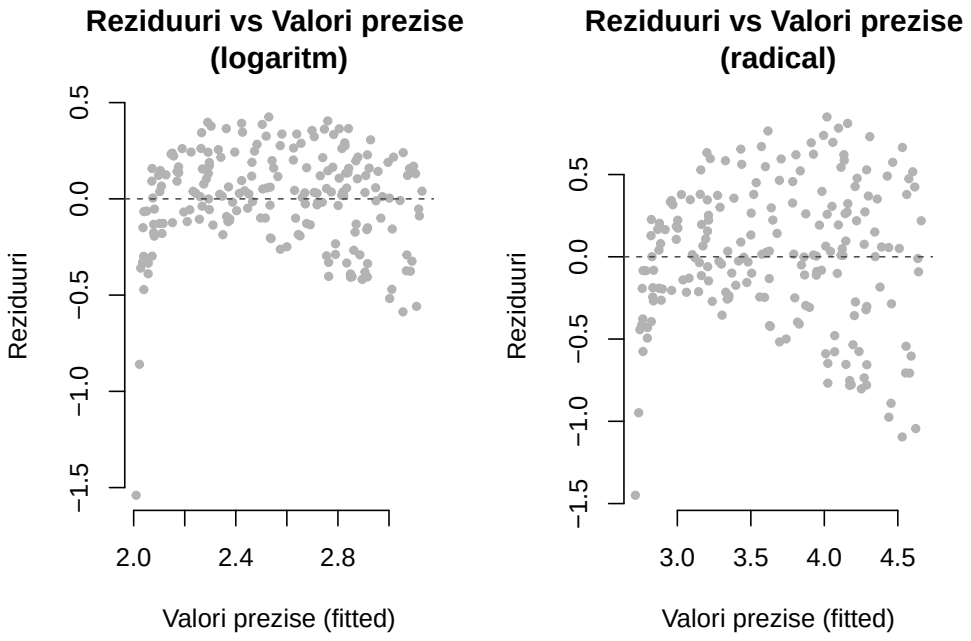
```
plot(residuals(advertise_TV_model)~fitted(advertise_TV_model),
     col = "grey70", pch = 16,
     xlab = "Valori prezise (fitted)",
     ylab = "Reziduuri",
     main = "Reziduuri vs Valori prezise",
```

```
     bty = "n")

abline(h = 0, col = "grey30")
```



Se poate observa că magnitudinea valorilor reziduale crește odată cu magnitudinea valorilor prezise (graficul are o formă de pâlnie) prin urmare ipoteza de homoscedasticitate nu este satisfăcută. O soluție în acest caz este de a transforma variabila răspuns Y cu ajutorul unei funcții concave, i.e. $\log Y$ sau $\sqrt{Y}$.

```
advertise_TV_model_log = lm(log(sales) ~ TV, data = advertise)
advertise_TV_model_sqrt = lm(sqrt(sales) ~ TV, data = advertise)

par(mfrow = c(1, 2))

plot(residuals(advertise_TV_model_log)~fitted(advertise_TV_model_log),
     col = "grey70", pch = 20,
     xlab = "Valori prezise (fitted)",
     ylab = "Reziduuri",
     main = "Reziduuri vs Valori prezise \n(logaritm)",
     bty = "n")

abline(h = 0, col = "grey30", lty = 2)

plot(residuals(advertise_TV_model_sqrt)~fitted(advertise_TV_model_sqrt),
     col = "grey70", pch = 20,
     xlab = "Valori prezise (fitted)",
     ylab = "Reziduuri",
     main = "Reziduuri vs Valori prezise \n(radical)",
     bty = "n")

abline(h = 0, col = "grey30", lty = 2)
```



Observăm că în cazul transformării logaritmice ($\log Y \approx \beta_0 + \beta_1 x + \varepsilon$) reziduurile par să îndeplinească ipoteza de homoscedasticitate (au variantă constantă). Cu toate acestea, graficul prezintă dovezi ale unei relații neliniare între volumul vânzărilor și bugetul alocat publicității TV.

Referințe

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2001. (Citat la pagina 19.)