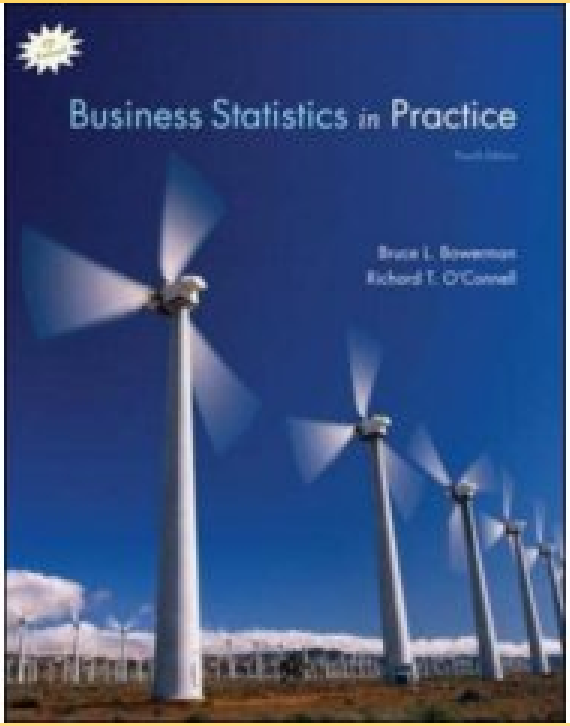
Analiza e regresionit të thjeshtë linear



Kapitulli 11

Analiza e regresionit të thjeshtë linear

Regresioni i thjeshtë linear

11.1 Modeli i regresionit të thjeshtë linear

11.2 Vlerësimet pikësore të katrorëve më të vegjël

11.3 Supozimet e modeleve dhe gabimi standard

11.4 Testimi i rëndësisë së pjerrtësisë dhe *y*-pikëprerjes

11.5 Intervalet e besueshmërisë dhe intervalet e parashikimit

11.6 Koeficienti i përcaktueshmërisë dhe korrelacionit

11.7 Testimi i rëndësisë së koeficientit të korelacionit të popullimit

11.1 Modeli i regresionit të thjeshtë linear

- ***Variabla e varur***, (ose ***reaguese***) është variabël me interes të cilën dëshirojmë ta kuptojmë ose parashikojmë.
- ***Variabla e pavarur***, (ose ***parashikuese***) është variabla të cilën do ta shfrytëzojmë për ta kuptuar ose parashikuar variablën e varur.
- ***Analiza e regresionit*** është një teknikë statistikore e cila shfrytëzon të dhëna të observuara për të vënë në relacion variablën e varur me një ose më tepër variabla të pavarura.

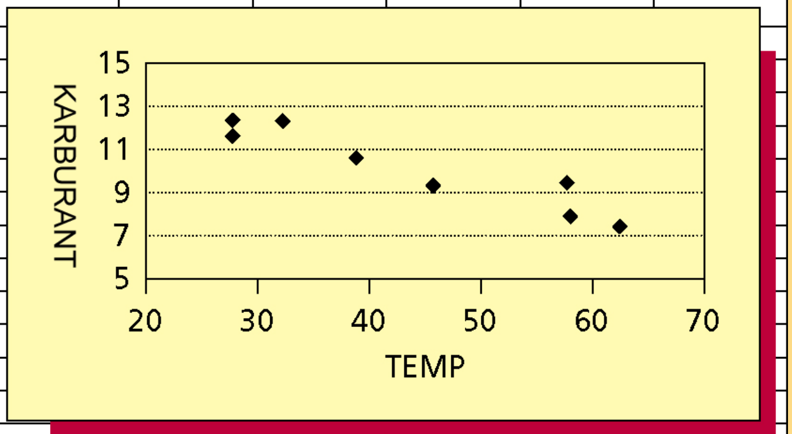
Modeli i regresionit të thjeshtë linear. (Vazhdim)

- Objektiva e analizës së regresionit është ndërtimi i një ***modeli regresioni*** (ose ekuacioni parashikues) i cili mund të shfrytëzohet për të përshkruar, parashikuar ose kontrolluar variablën e varur mbështetur në variablën e pavarur.
- ***Modeli i regresioni të thjeshtë linear*** supozon se relacioni ndërmjet variablës së varur y dhe variablës së pavarur x mund të përafrohet me anë të një drejtëze.
  - Më saktësisht, relacioni ndërmjet vlerës mesatare $\mu_{y|x}$ të variablës së varur y dhe variablës së pavarur x është linear.

Forma e modelit të regresionit të thjeshtë linear

$$y = \mu_{y|x} + \varepsilon \;= \beta_0 + \beta_1 x + \varepsilon$$

	A	B	C	D	E	F	G	H
1	TEMP	KARBURANT						
2	28	12.4						
3	28	11.7						
4	32.5	12.4						
5	39	10.8						
6	45.9	9.4						
7	57.8	9.5						
8	58.1	8						
9	62.5	7.5						
10								
11								
12								
13								
14								



Java	Temperatura mesatare për orë x (ºF)	Kosnumi javor i karburantit y (MMcf)
1	28.0	12.4
2	28.0	11.7
3	32.5	12.4
4	39.0	10.8
5	45.9	9.4
6	57.8	9.5
7	58.1	8.0
8	62.5	7.5

$\mu_{y|x} = \beta_0 + \beta_1 x$ është **vlera mesatare** e variablës së varur y kur vlera e variablës së pavarur është x.

β_0 është ***y*-pikëprerja**, mesatarja e y kur x është 0.

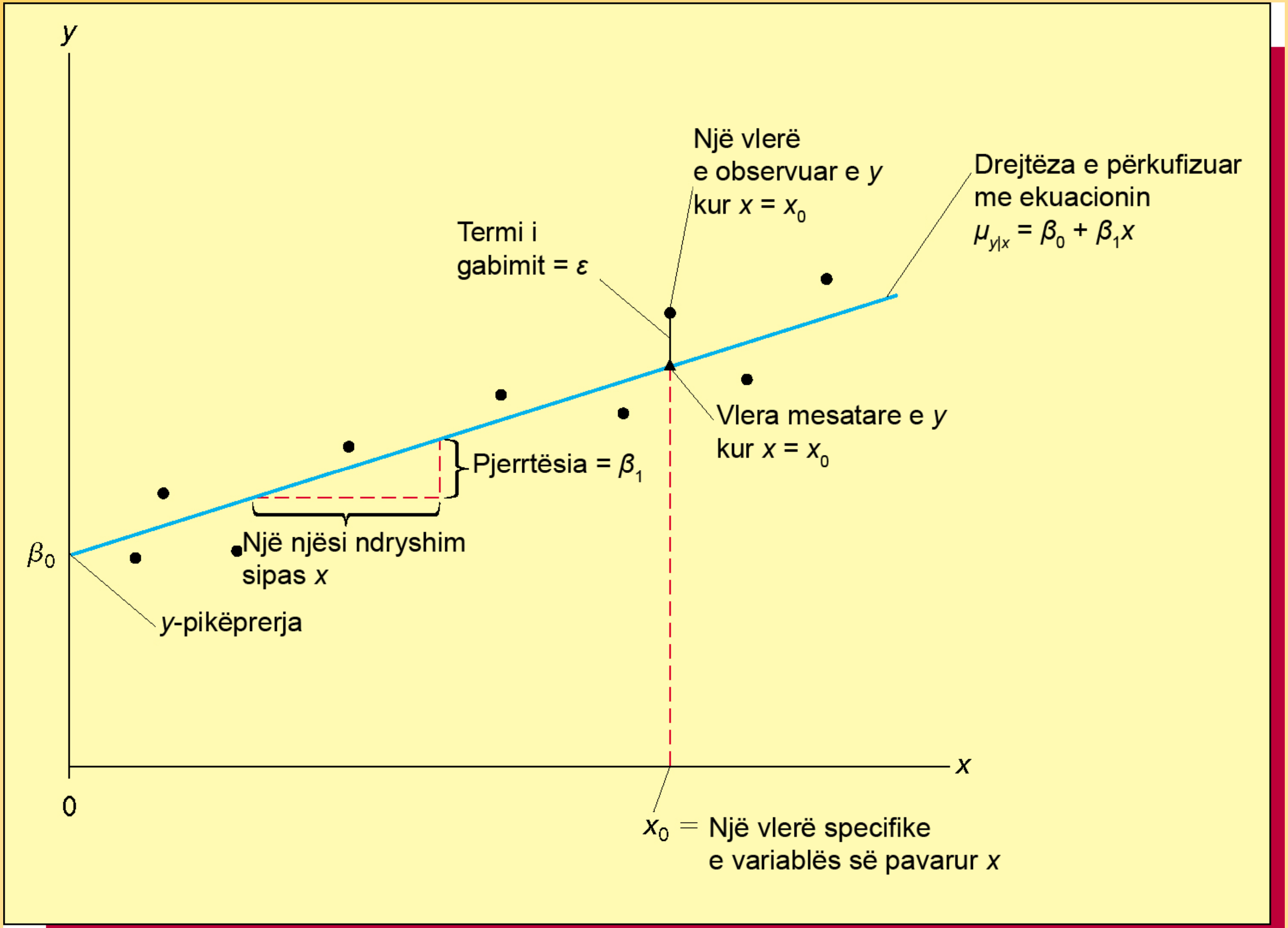
β_1 është **pjerrtësia**, mesatarja e ndryshimit të y për njësi ndryshimi të x.

ε është term **gabimi** që përshkruan efektin në y të të gjithë faktorëve të tjerë përveç x.

Termat e regresionit

- β_0 dhe β_1 quhen parametrat e regresionit
- β_0 është y-pikëprerja dhe β_1 është pjerrtësia
- Nuk janë të njohura vlerat e sakta të tyre
- Prandaj, duhet përdorur të dhëna mostre për ta përafruar vlerën e tyre (në pikën e ardhshme)

Ilustrimi i modelit të regresionit të thjeshtë linear



11.2 Vlerësimet pikësore të katrorëve më të vegjël

Ekuacioni parashikues: $\hat{y} = b_0 + b_1 x$

Vlerësimi i katrorëve më të vegjël për pjerrtësinë β_1:

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \qquad SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

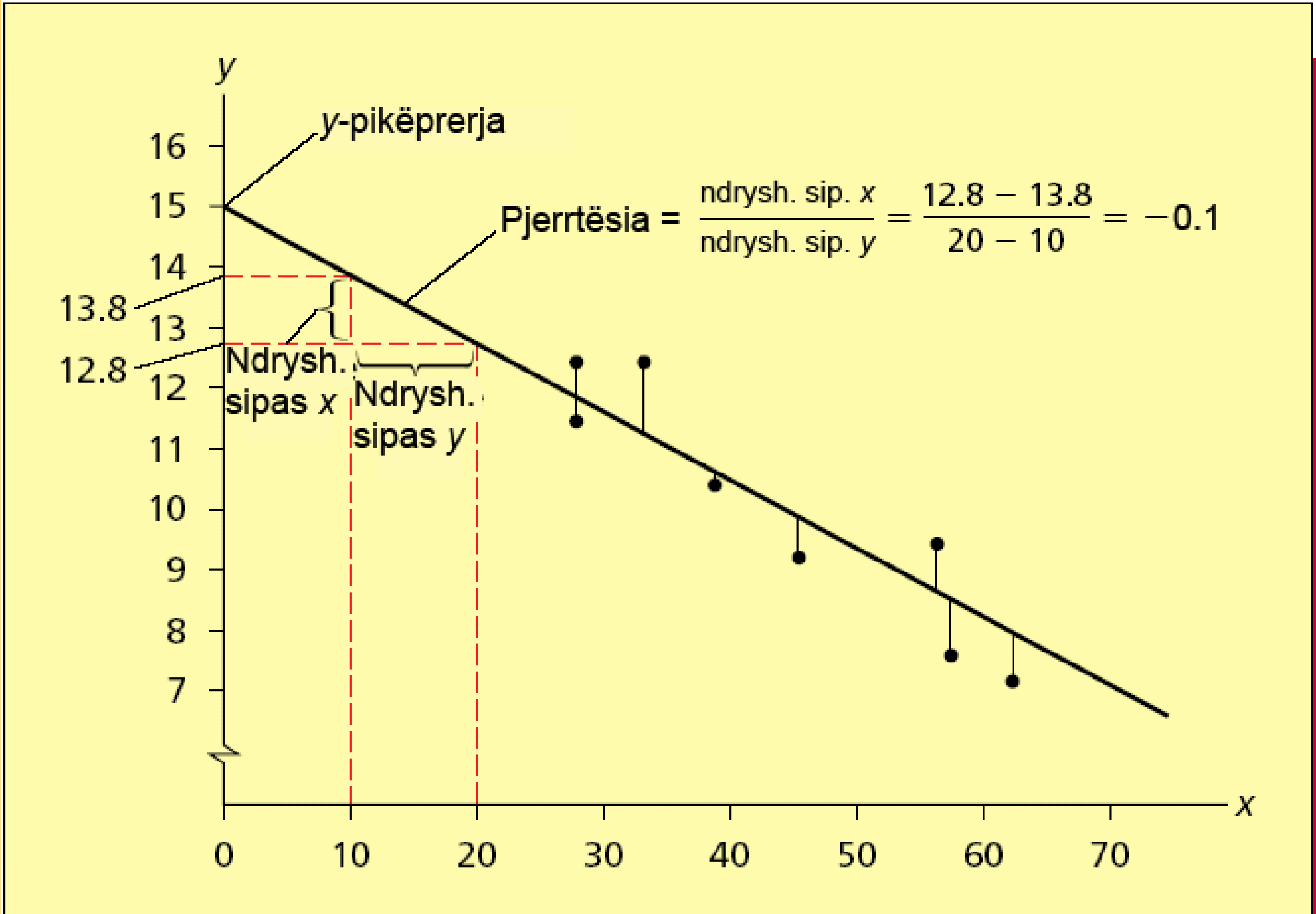
$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Vlerësimi i katrorëve më të vegjël për *y*-pikëprerjen β_0 :

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \bar{y} = \frac{\sum y_i}{n} \qquad \bar{x} = \frac{\sum x_i}{n}$$

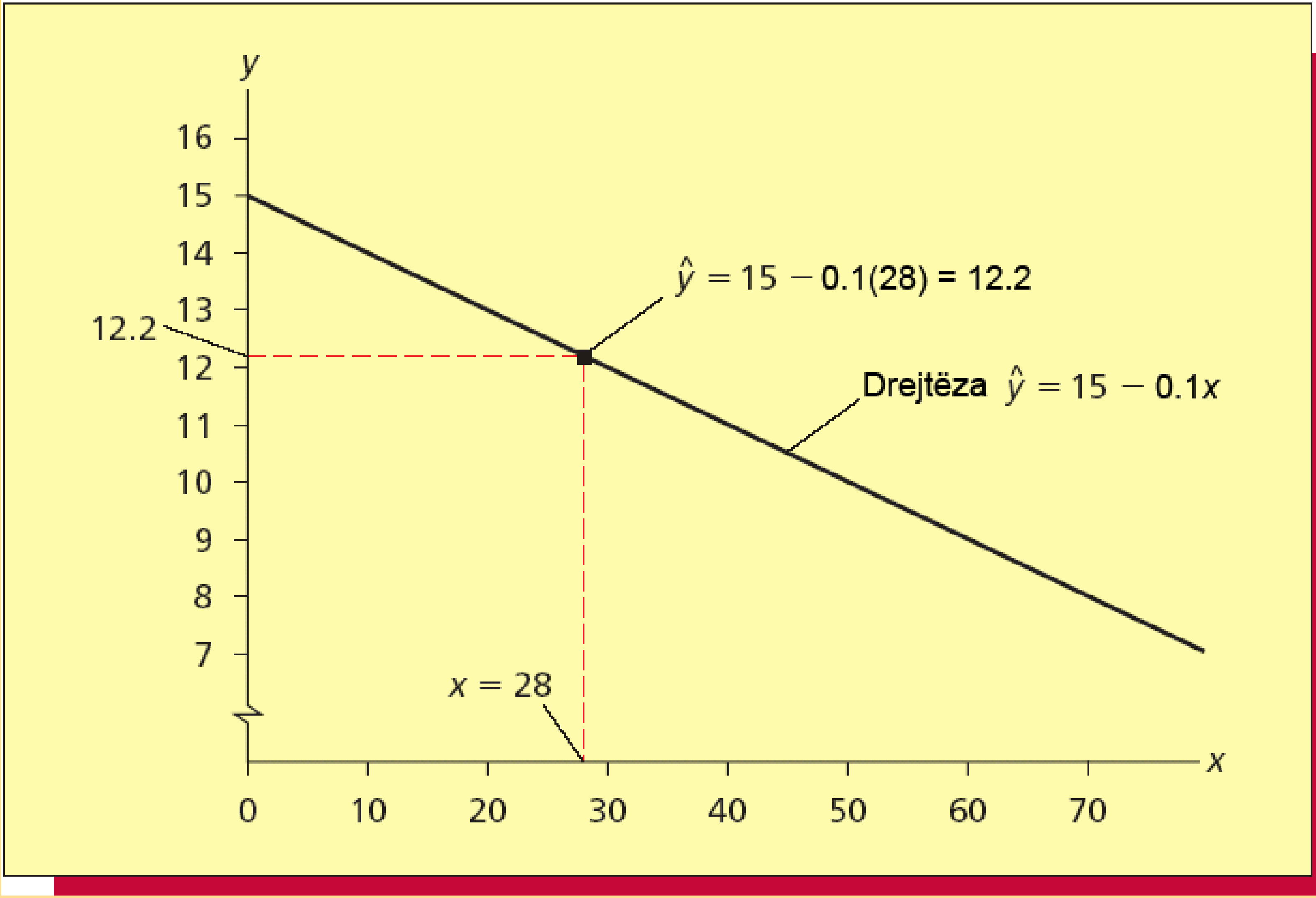
Shembull 11.3: Konsumi i karburantit

Futja vizuele e një drejtëze në të dhënat mbi konsumin e karburantit



Shembull 11.3: Konsumi i karburantit. (Vazhdim)

Shfrytëzimi i drejtëzës së futur në mënyrë vizuele për të parashikuar kur $x = 28$



Shembull 11.4: Konsumi i karburantit

y	x	x^2	xy
12.4	28.0	784.00	347.20
11.7	28.0	784.00	327.60
12.4	32.5	1056.25	403.00
10.8	39.0	1521.00	421.20
9.4	45.9	2106.81	431.46
9.5	57.8	3340.84	549.10
8.0	58.1	3375.61	464.80
7.5	62.5	3906.25	468.75
81.7	**351.8**	**16874.76**	**3413.11**

Shembull 11.4: Konsumi i karburantit. (Vazhdim)

- Nga slajdi i fundit
  - $\Sigma y_{\text{i}} = 81.7$
  - $\Sigma x_{\text{i}} = 351.8$
  - $\Sigma x^2_i = 16{,}874.76$
  - $\Sigma x_i y_i = 3{,}413.11$
- Pasi të kemi llogaritur një herë këto vlera, nuk kemi më tutje nevojë për të dhënat e papërpunuara
- Llogaritjet e b_0 dhe b_1 shfrytëzojnë këto totale

Shembull 11.4: Konsumi i karburantit. (Vazhdim)

- Pjerrtësia b_1

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$= 3413.11 - \frac{(351.8)(81.7)}{8} = -179.6475$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= 16874.76 - \frac{(351.8)^2}{8} = 1404.355$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-179.6475}{1404.355} = -0.1279$$

Shembull 11.4: Konsumi i karburantit. (Vazhdim)

- *y*-pikëprerja b_0

$$\bar{y} = \frac{\sum y_i}{n} = \frac{81.7}{8} = 10.2125$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{351.8}{8} = 43.98$$

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ &= 10.2125 - (-0.1279)(43.98) \\ &= 15.84 \end{aligned}$$

Shembull 11.4: Konsumi i karburantit. (Vazhdim)

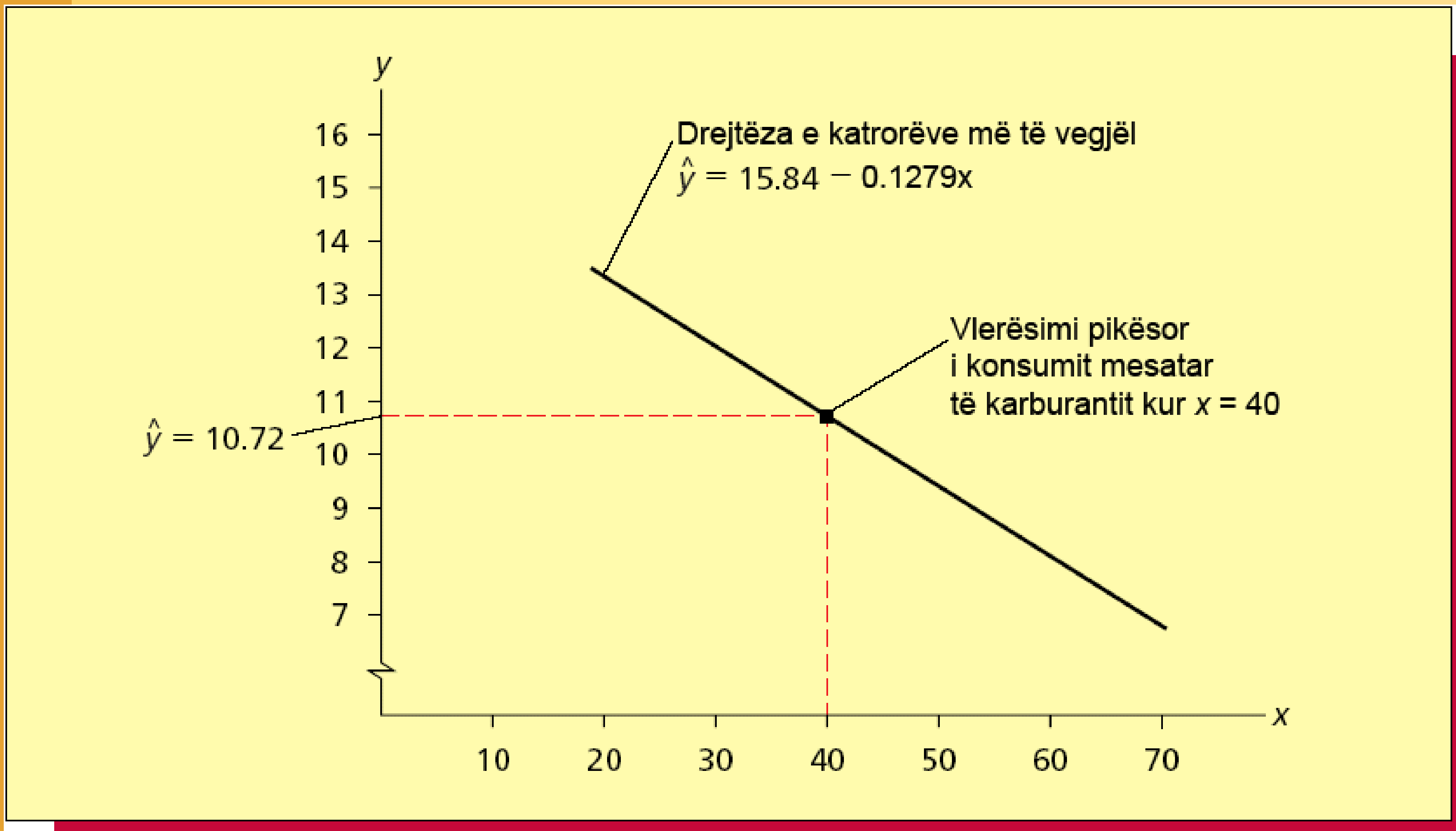
- Parashikimi ($x = 40$)

$$\hat{y} = b_0 + b_1 x$$

$$= 15.84 - 0.1279(40)$$

$$= 10.72 \text{ MMcf karburant}$$

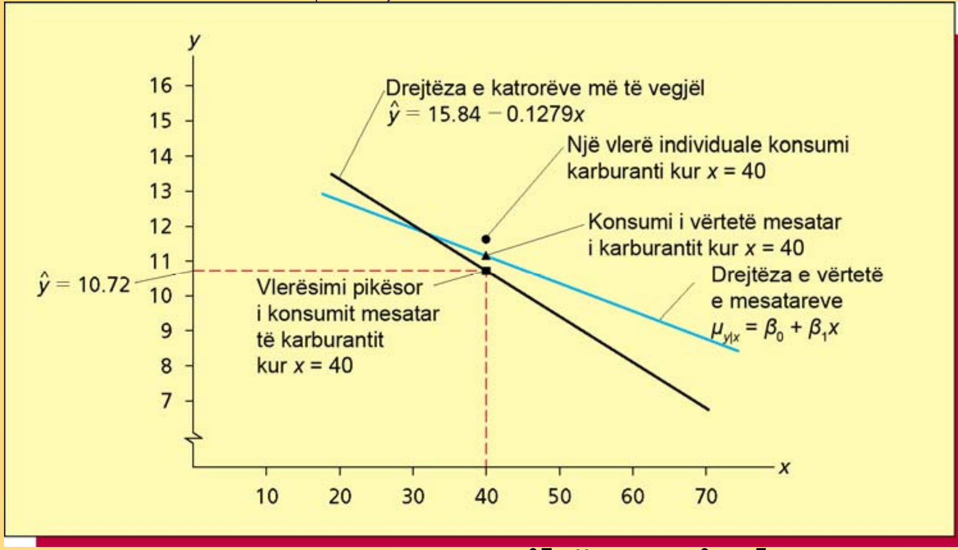
Shembull 11.4: Konsumi i karburantit. (Vazhdim)



Shembujt e konsumit të karburantit

Parashikimi $(x = 40)$ $\hat{y} = b_0 + b_1 x = 15.84\text{-}0.1279(40) = 10.72$ MMcf karburanti

y	x	x^2	xy
12.4	28.0	784.00	347.20
11.7	28.0	784.00	327.60
12.4	32.5	1056.25	403.00
10.8	39.0	1521.00	421.20
9.4	45.9	2106.81	431.46
9.5	57.8	3340.84	549.10
8.0	58.1	3375.61	464.80
7.5	62.5	3906.25	468.75
81.7	**351.8**	**16874.76**	**3413.11**



Pjerrtësia b_1

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 3413.11 - \frac{(351.8)(81.7)}{8} = -179.6475$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 16874.76 - \frac{(351.8)^2}{8} = 1404.355$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-179.6475}{1404.355} = \mathbf{-0.1279}$$

***y*-pikëprerja b_0**

$$\bar{y} = \frac{\sum y_i}{n} = \frac{81.7}{8} = 10.2125$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{351.8}{8} = 43.98$$

$$b_0 = \bar{y} - b_1\bar{x} = 10.2125 - (-0.1279)(43.98) = \mathbf{15.84}$$

11.3 Supozimet e modelit të regresionit

Modeli $y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$

Supozimet mbi termat e gabimit të modelit, ε-ët

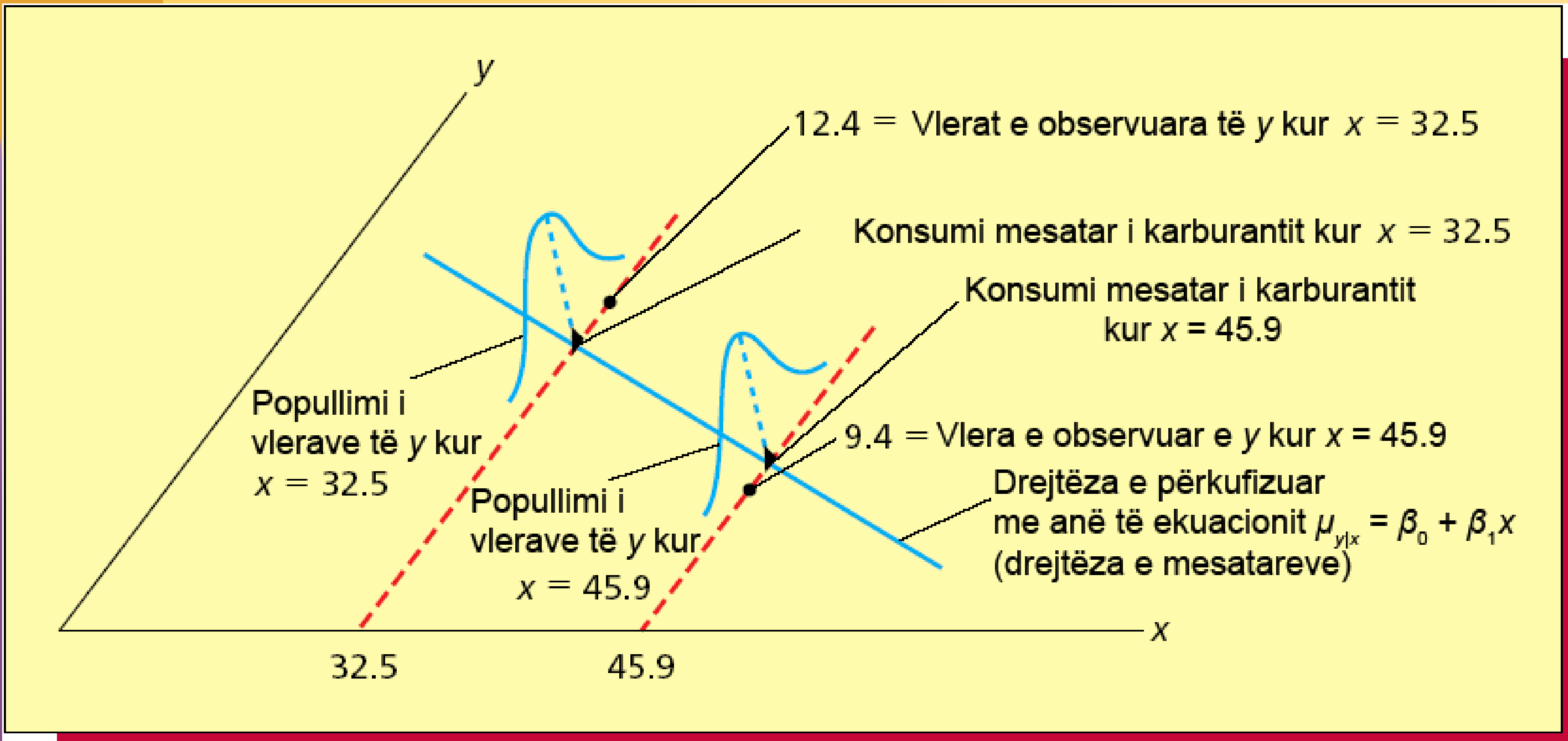
1. **Mesatarja zero:** Për çdo vlerë të dhënë x, popullimi i vlerave potenciale të termave të gabimit ka mesataren 0.

2. **Varianca konstante:** Për çdo vlerë të dhënë x, popullimi i vlerave potenciale të termave të gabimit ka variancë e cila nuk varet nga vlera e x.

3. **Normaliteti:** Për çdo vlerë të dhënë x, popullimi i vlerave potenciale të termave të gabimit ka shpërndarje normale.

4. **Pavarësia:** Vlerat e termave të gabimit janë statistikisht të pavarur nga njëri tjetri.

Supozimet e modelit të regresionit të ilustruara



Shuma e katrorëve të gabimeve

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Katrori mesatar i gabimit

- Është vlerësim pikësor i variancës së gabimit (së mbetjeve)
- *SSE* është nga slajdi paraprak

$$s^2 = MSE = \frac{SSE}{n-2}$$

Gabimi standard

- Është vlerësim pikësor i devijimit standard të gabimit (të mbetjeve)
- *MSE* është nga slajdi paraprak

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

Katrori mesatar i gabimit dhe gabimi standard

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \qquad s^2 = MSE = \frac{SSE}{n-2} \qquad s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

Shembull 11.6 Rasti i konsumit të karburantit

y	x	parash	y - parash	(y - parash)^2
12.4	28.0	12.2588	0.1412	0.019937
11.7	28.0	12.2588	-0.5588	0.312257
12.4	32.5	11.6833	0.7168	0.513731
10.8	39.0	10.8519	-0.0519	0.002694
9.4	45.9	9.9694	-0.5694	0.324205
9.5	57.8	8.4474	1.0526	1.108009
8.0	58.1	8.4090	-0.4090	0.167289
7.5	62.5	7.8463	-0.3462	0.119889
			SSE	**2.568011**

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{2.568}{6} = \mathbf{0.428}$$

$$s = \sqrt{s^2} = \sqrt{0.428} = \mathbf{0.6542}$$

11.4 Testi i rëndësisë dhe vlerësimi për pjerrtësinë

- Një model regresioni nuk ka të ngjarë të jetë i dobishëm përveç në qoftë se ekziston relacion i rëndësishëm ndërmjet x dhe y
- Për të testuar rëndësinë, shfrytëzojmë zero hipotezën:
  $H_0: \beta_1 = 0$
- Kundrejt hipotezës alternative:
  $H_a: \beta_1 \neq 0$

11.4 Testi i rëndësisë dhe vlerësimi për pjerrtësinë

Në qoftë se vlejnë supozimet e regresionit, mund **hedhim poshtë H_0: $\beta_1 = 0$ me nivel rëndësie α** atëherë dhe vetëm atëherë kur vlen **rregulla e hedhjes poshtë** përkatëse ose, ekuivalente me të, në qoftë se *p*-vlera përkatëse është më e vogël se α.

Alternative	**Hedh poshtë H_0 nëse:**	***p*-vlera**
$H_a: \beta_1 > 0$	$t > t_\alpha$	Bishti i djathtë nën *t*-shpërndarjen i dhënë me *t*
$H_a: \beta_1 < 0$	$t < -t_\alpha$	Bishti i majtë nën *t*-shpërndarjen i dhënë me *t*
$H_a: \beta_1 \neq 0$	$\lvert t \rvert > t_{\alpha/2}$, that is $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$	Bishti i dyfishtë nën *t*-shpërndarjen i dhënë me $\lvert t \rvert$

Statistika e testit

$$t = \frac{b_1}{s_{b_1}} \quad \text{ku} \quad s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

100(1-α)% intervali i besueshmërisë për β_1

$$[b_1 \pm t_{\alpha/2} s_{b_1}]$$

t_α, $t_{\alpha/2}$ dhe *p*-vlerat mbështetën në $n - 2$ shkallë lirie.

11.4 Testi i rëndësisë dhe vlerësimi për *y*-pikëprerjen

Në qoftë se vlejnë supozimet e regresionit, mund **hedhim poshtë H_0: $\beta_0 = 0$ me nivel rëndësie α** atëherë dhe vetëm atëherë kur vlen **rregulla e hedhjes poshtë** përkatëse ose, ekuivalente me të, në qoftë se *p*-vlera përkatëse është më e vogël se α.

Alternative	**Hedh poshtë H_0 nëse:**	***p*-vlera**
$H_a: \beta_0 > 0$	$t > t_\alpha$	Bishti i djathtë nën *t*-shpërndarjen i dhënë me *t*
$H_a: \beta_0 < 0$	$t < -t_\alpha$	Bishti i majtë nën *t*-shpërndarjen i dhënë me *t*
$H_a: \beta_0 \neq 0$	$\lvert t \rvert > t_{\alpha/2}$, that is $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$	Bishti i dyfishtë nën *t*-shpërndarjen i dhënë me |t|

Statistika e testit

$$t = \frac{b_0}{s_{b_0}} \quad \text{ku} \quad s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

100(1-α)% intervali i besueshmërisë për β_0

$$[b_0 \pm t_{\alpha/2} s_{b_0}]$$

t_α, $t_{\alpha/2}$ dhe *p*-vlerat mbështetën në $n - 2$ shkallë lirie.

11.5 Intervalet e besueshmërisë dhe të parashikimit

- Pika në drejtëzën e regresionit e cila i korrespondon një vlere të veçantë x_0 të variablës së pavarur x është

$$\hat{y} = b_0 + b_1 x_0$$

- Nuk ka të ngjarë që kjo vlerë të jetë e barabartë me vlerën mesatare të y kur $x = x_0$
- Prandaj na duhen kufij mbi distancën e vlerës së parashikuar nga vlera aktuale
- Llogarisim një interval besueshmërie për vlerën mesatare të y dhe një interval parashikimi për një vlerë individuale të y

Vlera e distancës

- Që të dytë: intervali i besueshmërisë për vlerën mesatare të y dhe intervali i parashikimit për një vlerë individuale të y kanë të bëjnë me vlerën e distancës
- ***Vlera e distancës*** për një vlerë të caktuar të x_0 të x është

$$\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}$$

- Vlera e distancës është masë për distancën ndërmjet vlërës x_0 të x dhe $\bar{x}$
- Sa më e madhe të jetë distanca e x_0 nga x, aq më e madhe është vlera e distancës

Intervali i besueshmërisë për vlerën mesatare të *y*

- Supozojmë se vlejnë supozimet e regresionit
- Formula për $100(1-\alpha)\%$ intervalin e besueshmërisë për vlerën mesatare të y është

$$[\hat{y} \pm t_{\alpha/2} s \sqrt{\text{Vlera e distancës}}]$$

- Mbështetet në $n-2$ shkallë lirie

Intervali i besueshmërisë për një vlerë individuale të *y*

- Supozojmë se vlejnë supozimet e regresionit
- Formula për $100(1 - \alpha)\%$ intervalin e besueshmërisë për vlerën mesatare të y është

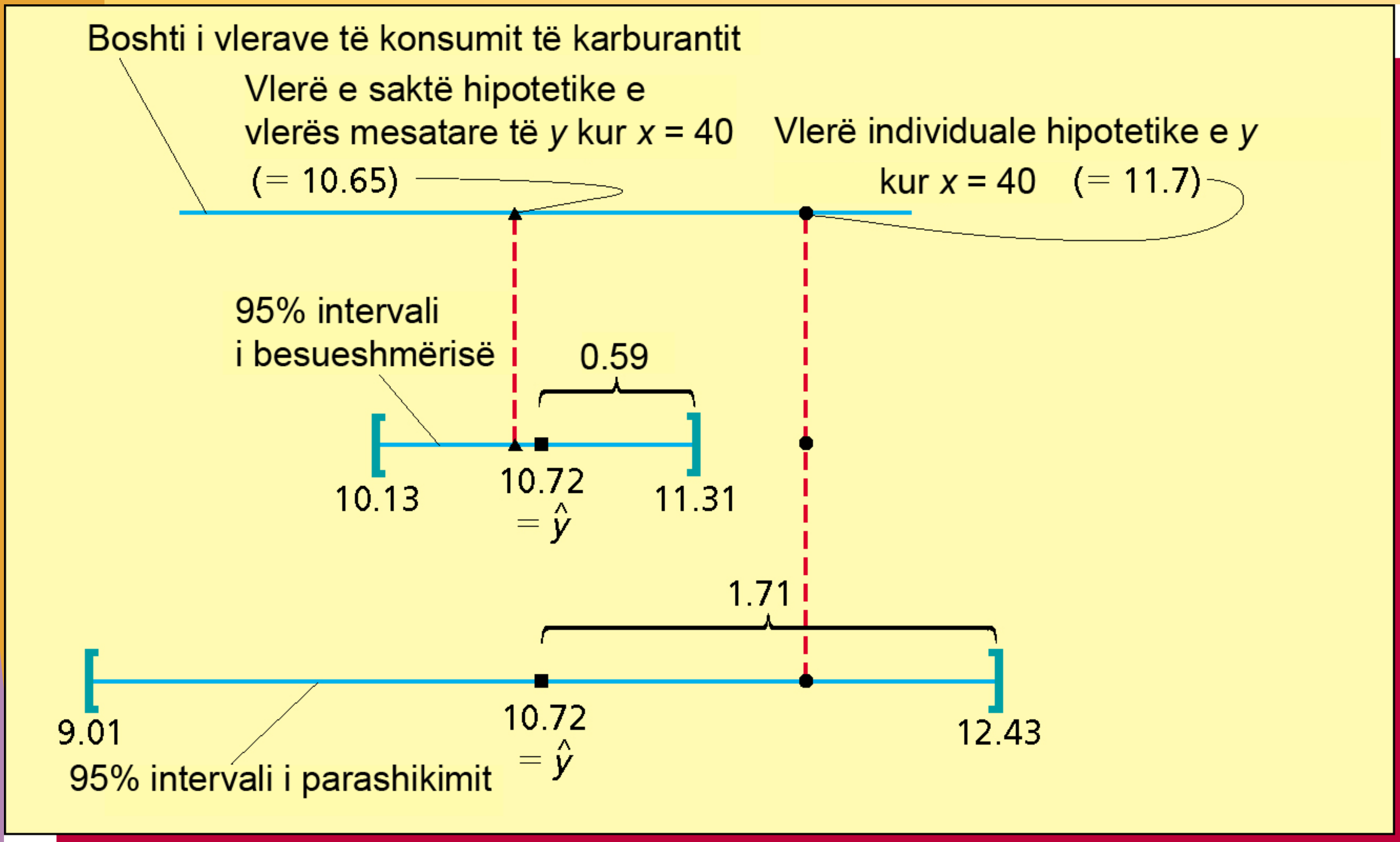
$$[\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \text{Vlera e distancës}}]$$

- Mbështetet në $n - 2$ shkallë lirie

Cili të përdoret?

- Intervali i parashikimit është i dobishëm në qoftë se është me rëndësi të prashikohet një vlerë individuale e variablës së varur.
- Intervali i besueshmërisë është i dobishëm në qoftë se është me rëndësi të vlerësohet mesatarja.
- Intervali i parashikimit është gjithmonë më i gjerë sesa intervali i besueshmërisë.

Shembuj 11.10-11: Rasti i konsumit të karburantit



11.6 Koeficienti i thjeshtë i përcaktueshmërisë dhe korrelacionit

- Sa i dobishëm është një model i caktuar regresioni?
- Një masë dobishmërie është koeficienti i thjeshtë i përcaktueshmërisë
- Shënohet me simbolin r^2

Llogaritja e koeficientit të thjeshtë të përcaktueshmërisë

1. Variacioni total:

$$\sum (y_i - \bar{y})^2$$

2. Variacioni i shpjeguar:

$$\sum (\hat{y}_i - \bar{y})^2$$

3. Variacioni i pashpjeguar:

$$\sum (y_i - \hat{y}_i)^2$$

4. Variacioni total:

$$\text{Variacioni total} = \text{Variacioni i shpjeguar} + \text{Variacioni i pashpjeguar}$$

5. Koeficienti i thjeshtë i përcaktueshmërisë

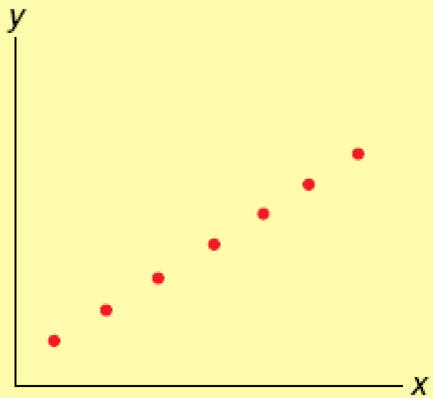
$$r^2 = \frac{\text{Variacioni i shpjeguar}}{\text{Variacioni total}}$$

Koeficienti i thjeshtë i korrelacionit

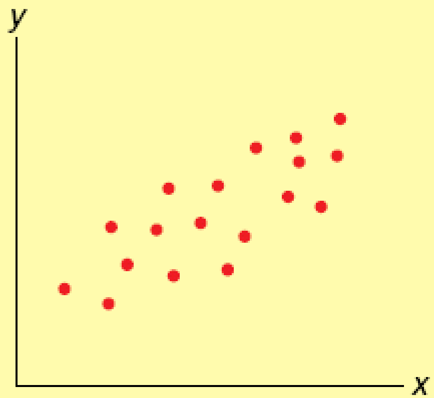
- Koeficienti i thjeshtë i korrelacionit mat fuqinë e relacionit linear ndërmjet *y* dhe *x*, dhe shënohet me *r*

$$r = +\sqrt{r^2} \text{ në qoftë se } b_1 \text{ është pozitiv, dhe}$$
$$r = -\sqrt{r^2} \text{ në qoftë se } b_1 \text{ është negativ}$$

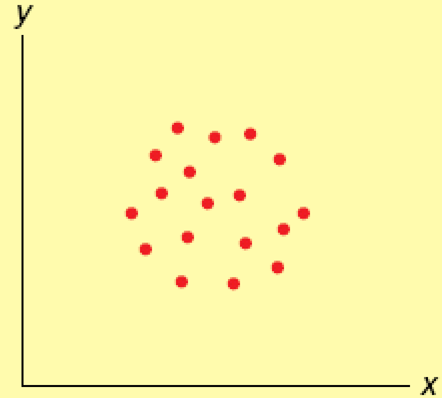
Vlera të ndryshme të koeficientit të korrelacionit



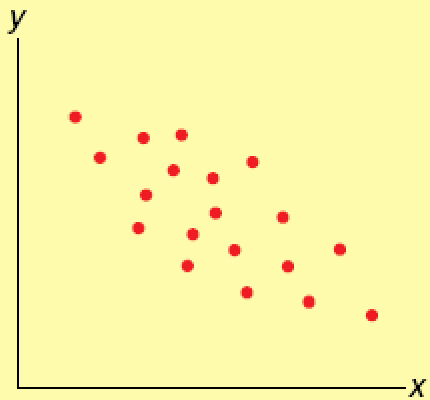
(a) r = 1: korrelacion ideal pozitiv



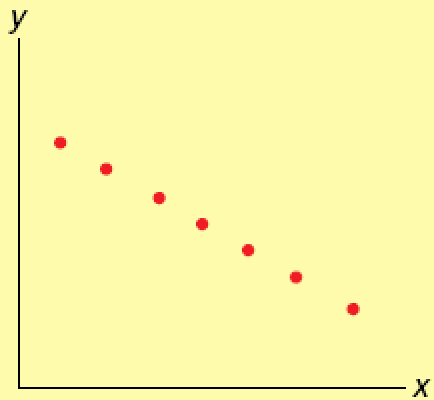
(b) Korrelacion pozitiv ($r > 0$): *y* rritet me rritjen e *x* në mënyrë drejtvizore



(c) Pak korrelacion (*r* afër 0): pak relacion linear ndërmjet *y* dhe *x*



(d) Korrelacion negativ ($r < 0$): *y* zvogëlohet me rritjen e *x* në mënyrë drejtvizore



(e) r = 1: korrelacion ideal negativ

Testimi i rëndësisë së koeficientit të korrelacionit të popullimit

- Koeficienti i thjeshtë i korrelacionit (r) mat relacionin linear ndërmjet vlerave të observuara të x dhe vlerave të observuara të y nga mostra
- Koeficienti i korrelacionit të popullimit (ρ) mat relacionin linear ndërmjet të gjitha kombinimeve të mundura të vlerave të observuara të x dhe y
- r është përafrim i ρ

Testimi i ρ

- Testimi se a është korrelacioni i rëndësishëm
  $H_0: \rho = 0$
  $H_a: \rho \neq 0$
- Statistika e testit

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

- Rezultat të njëjtë sikur testi për β_1

Regresioni i thjeshtë linear

Përfundim:

11.1 Modeli i regresionit të thjeshtë linear

11.2 Vlerësimet pikësore të katrorëve më të vegjël

11.3 Supozimet e modeleve dhe gabimi standard

11.4 Testimi i rëndësisë së pjerrtësisë dhe *y*-pikëprerjes

11.5 Intervalet e besueshmërisë dhe intervalet e parashikimit

11.6 Koeficienti i përcaktueshmërisë dhe korrelacionit

11.7 Testimi i rëndësisë së koeficientit të korelacionit të popullimit