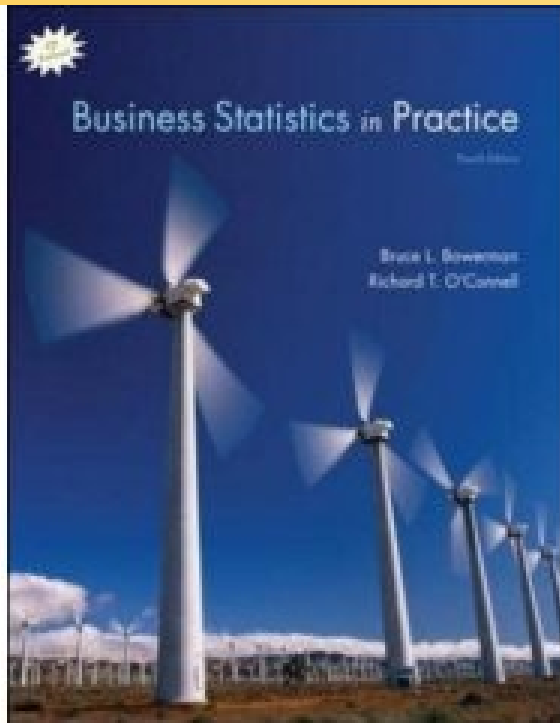


Regresioni i shumëfishtë dhe ndërtimi i modeleve



Kapitulli 15

Regresioni i shumëfishtë
dhe ndërtimi i modeleve

Regresioni i shumëfishtë dhe ndërtimi i modeleve

- 15.1 Modeli i regresionit të shumëfishtë dhe vlerësimi pikësor i katrorëve më të vegjël
- 15.2 Supozimet e modeleve dhe gabimi standard
- 15.3 R^2 dhe R^2 i përshtatur
- 15.4 F -testi i përgjithshëm
- 15.5 Testimi i rëndësisë së një ndryshoreje të pavarur
- 15.6 Intervallet e konfidencës dhe ato të parashikimit
- 15.7 Rasti i performimit të teritorit të shitjes

Regresioni i shumëfishtë dhe ndërtimi i modeleve (Vazhdim)

- 15.8 Përdorimi i dummy-variablave për modelimin e variablave të pavarura kualitative
- 15.9 Përdorimi i variancave katrore dhe atyre të interaksionit
- 15.10 Ndërtimi i modeleve dhe efekti i multikolinearitetit
- 15.11 Analiza reziduale në regresionin e shumëfishtë
- 15.12 Regresioni logjistik

15.1 Modeli i regresionit të shumëfishtë dhe vlarësimi pikësor me katrorët më të vegjël

- Regresioni i thjeshtë linear përdorte një variabël të pavarur për të shpjeguar variablën e varur.
 - Disa vartësi janë tepër komplekse për t'u përshkruar me një ndryshore të vetme të pavarur.
- Modelet e regresionit të shumëfishtë përdorin dy ose më tepër variabla të pavarura për të përshkruar variablën e varur.
 - Një gjë e tillë u lejon modeleve të regresionit të shumëfishtë të merren me situata më komplekse.
 - Nuk ka ndonjë kufi të numrit të variablave të pavarura të cilat mund t'i përdorë një model.
- Regresioni i shumëfishtë ka vetëm një variabël të varur.

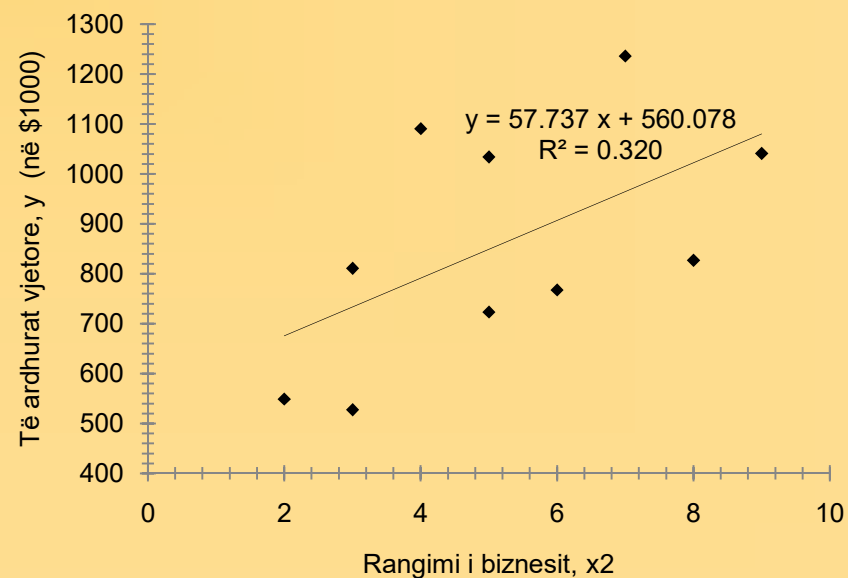
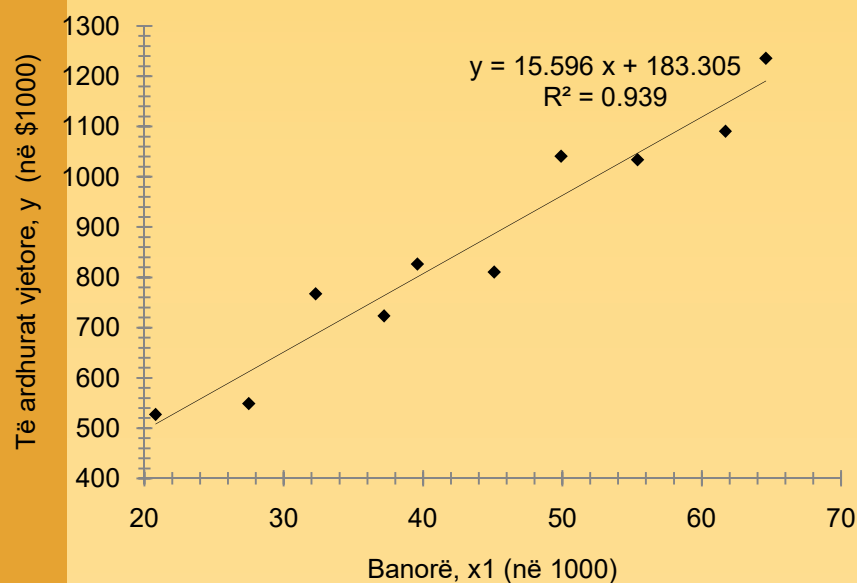
Modeli i regresionit linear

- *Modeli i regresionit linear* i cili lidh y me x_1, x_2, \dots, x_k është
$$y = \mu_{y|x_1, x_2, \dots, x_k} + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$
ku
- $\mu_{y|x_1, x_2, \dots, x_k} + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$ është *vlera mesatare* e ndryshores së varur y kur vlerat e ndryshoreve të pavarura janë x_1, x_2, \dots, x_k .
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ janë parametrat e regresionit që lidhin vlerën mesatare të y me x_1, x_2, \dots, x_k .
- ε është termi i *gabimit* që përshkruan efektin në y të të gjithë faktorëve tjerë përveç variablave të pavarura x_1, x_2, \dots, x_k .

Shembull: Modeli i regresionit linear

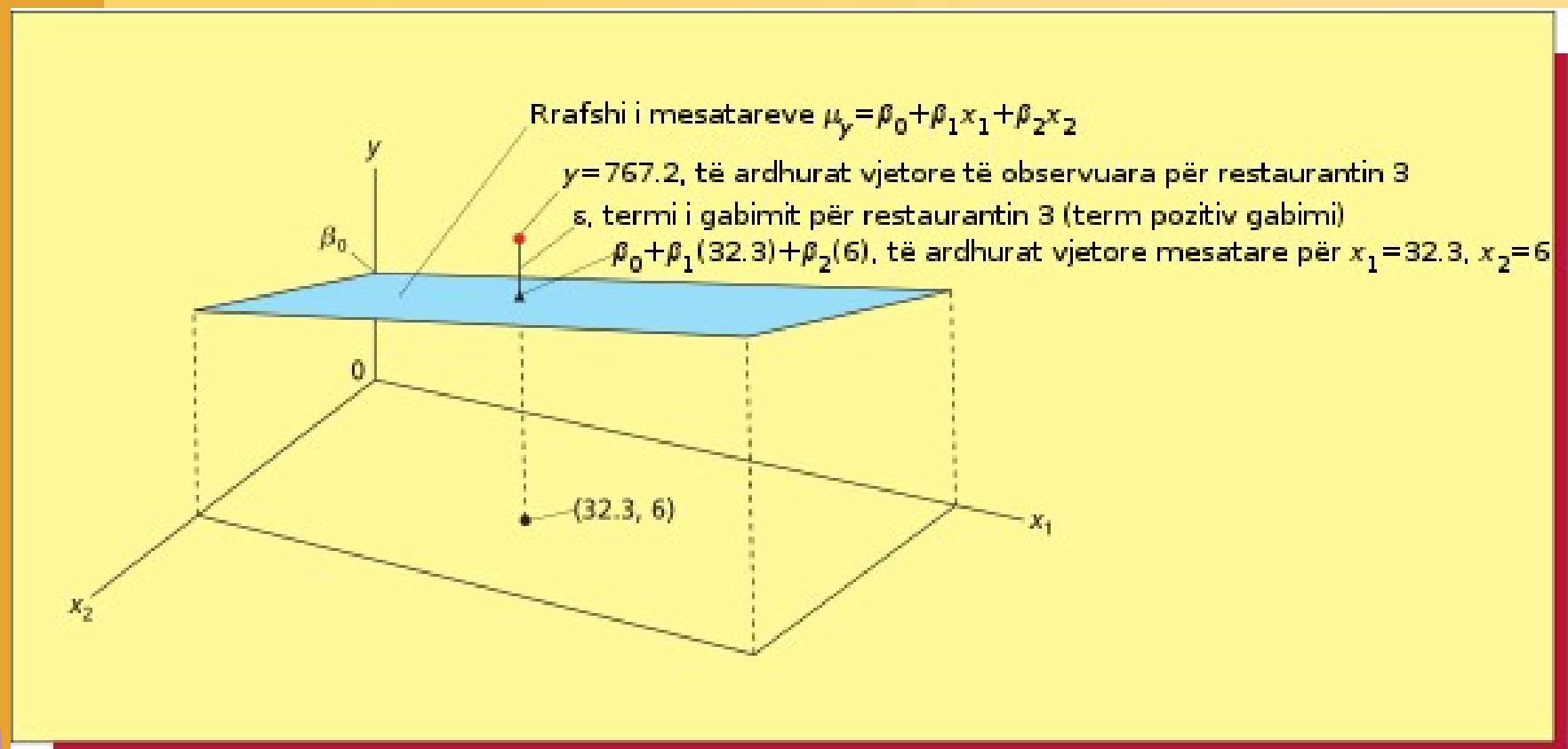
- Shembull 15.1. Rasti Tasty Sub

Restauranti	Banorë, x1 (1000)	Rangimi i biznesit, x2	Të ardhurat vjetore, y (\$1000)
1	20.8	3	527.1
2	27.5	2	548.7
3	32.3	6	767.2
4	37.2	5	722.9
5	39.6	8	826.3
6	45.1	3	810.5
7	49.9	9	1040.7
8	55.4	5	1033.6
9	61.7	4	1090.3
10	64.6	7	1235.8



Ilustrimi i modelit të regresionit që lidh y me x_1 dhe x_2

Shembulli 15.1. Rasti Tasty Sub



Vlerësimi pikësor i katrorëve më të vegjël

- *Ekuacioni i vlerësimit/parashikimit:*

$$\hat{y} = b_0 + b_1 x_{01} + b_2 x_{02} + \dots + b_k x_{0k}$$

është **vlerësim pikësor i vlerës mesatare** të variablës së varur y kur vlerat e variablave të pavarura janë $x_{01}, x_{02}, \dots, x_{0k}$.

- Është poashtu **parashikimi pikësor i një vlere individuale** të variablës së varur y kur vlerat e variablave të pavarura janë $x_{01}, x_{02}, \dots, x_{0k}$.
- b_1, b_2, \dots, b_k janë përafrime të katrorëve më të vegjël të parametrave $\beta_1, \beta_2, \dots, \beta_k$
- $x_{01}, x_{02}, \dots, x_{0k}$ janë vlera specifike të variablave të pavarura x_1, x_2, \dots, x_k .

15.2. Supozimet e modelit dhe gabimi standard

- Modeli

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

- Supozimet mbi modelin e regresionit të shumfishtë formulohen sipas termit të gabimit ε :
 - **Mesatarja zero:** Mesatarja e termit të gabimit ε është 0
 - **Varianca konstante:** Varianca e termit të gabimit ε nuk varet nga kombinimi i vlerave të x_1, x_2, \dots, x_k
 - **Normaliteti:** Termi i gabimit ε ka shpërndarje normale për çfarëdo kombinimi të vlerave të x_1, x_2, \dots, x_k
 - **Pavarësia:** Vlerat e termit të gabimit janë statistikisht të pavarura nga njëra tjetra

Gabimi mesatar katror dhe gabimi standard

- *Shuma e katrorëve të gabimeve (Sum of Squared Errors):*

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

- *Gabimi mesatar katror (Mean Square Error):* vlerësim pikësor i variancës σ^2 të mbetjeve (gabimeve)

$$s^2 = MSE = \frac{SSE}{n - (k + 1)}$$

- *Gabim standard:* Vlerësim pikësor i devijimit standard σ mbetjeve (gabimeve)

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n - (k + 1)}}$$

15.3. R^2 dhe R^2 i përshtatur

- *Variacioni total:*

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

- *Varicacioni i shpjeguar:*

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

- *Variacioni i pashpjeguar:*

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

- Variacioni total është shuma e variacionit të shpjeguar dhe atij të pashpjeguar.
- *Koeficienti i shumëfishtë i përcaktueshmërisë* R^2 është herësi ndërmjet variacionit të shpjeguar dhe variacionin total.
- R^2 është proporcioni i variacionit total i cili është shpjeguar nga modeli i regreionit të shumëfishtë.
- *Koeficienti i korrelacionit të shumëfishtë* është $R = \sqrt{R^2}$

R^2 i përshtatur

- *Koeficienti i përshtatur i shumëfishtë i përcaktueshmërisë është*

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1} \right) \left(\frac{n-1}{n-(k+1)} \right)$$

- Për të evituar mbivlerësimin e rëndësisë së variablave të pavarura, shumë analistë rekomandojnë përdorimin e koeficientit të përshtatur të shumfishtë të përcaktueshmërisë.

R^2 i përshtatur

- Shembulli 15.2. Rasti i Tasty Sub:

Regression Analysis						
	R ²	0.981				
	Adjusted R ²	0.976	n	10		
	R	0.990	k	2		
	Std. Error	36.686	Dep. Var.	Të ardhurat vjetore, y (në \$1000)		
ANOVA table						
Source	SS	df	MS	F	p-value	
Regression	486,355.6632	2	243,177.8316	180.69	9.46E-07	
Residual	9,420.8458	7	1,345.8351			
Total	495,776.5090	9				
Regression output						
					confidence interval	
variables	coefficients	std. error	t (df=7)	p-value	95% lower	95% upper
Intercept	125.2888	40.9333	3.061	.0183	28.4969	222.0807
Banorë, x1 (në 1000)	14.1996	0.9100	15.604	1.07E-06	12.0478	16.3515
Rangimi i biznesit, x2	22.8107	5.7692	3.954	.0055	9.1686	36.4527

- Modeli i regresionit të shumëfishtë:

$$\hat{y} = 125.2888 + 14.1996 x_{01} + 22.8107 x_{02}$$

15.4. F -testi i përgjithshëm

- Testohen:

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ kundrejt

H_a : Së paku njëra nga $\beta_1, \beta_2, \dots, \beta_k \neq 0$

- Statistika e testit është:

$$F(\text{model}) = \frac{(\text{Variacioni i shpjeguar})/k}{(\text{Variacioni i pashpjeguar})/(n - (k + 1))}$$

- Hedh poshtë H_0 në favor të H_a në qoftë se $F(\text{model}) > F_\alpha$ ose $p\text{-vlera} < \alpha$.
- F_α mbështetet në k mbi $n - (k+1)$ shkallë lirie.

Shembulli 15.3. Rasti i Tasty Sub

Regression Analysis							
		R²	0.981				
		Adjusted R²	0.976	n	10		
		R	0.990	k	2		
		Std. Error	36.686	Dep. Var.	Të ardhurat vjetore, y (në \$1000)		
ANOVA table							
	Source	SS	df	MS	F	p-value	
	Regression	486,355.6632	2	243,177.8316	180.69	9.46E-07	
	Residual	9,420.8458	7	1,345.8351			
	Total	495,776.5090	9				
Regression output							
	variables	coefficients	std. error	t (df=7)	p-value	confidence interval	
						95% lower	95% upper
	Intercept	125.2888	40.9333	3.061	.0183	28.4969	222.0807
	Banorë, x1 (në 1000)	14.1996	0.9100	15.604	1.07E-06	12.0478	16.3515
	Rangimi i biznesit, x2	22.8107	5.7692	3.954	.0055	9.1686	36.4527

15.5. Testimi i rëndësisë së një ndryshoreje të pavarur

- Një ndryshore në një model regresioni të shumëfishtë nuk ka shumë të ngjarë të jetë e dobishme përveç në qoftë se ka një lidhmëri të rëndësishme ndërmjet saj dhe y .
- Për të testuar rëndësinë e ndryshores x_j , përdorim hipotezën zero

$$H_0: \beta_j = 0$$

- Hedh poshtë H_0 në favor të

$$H_a: \beta_j \neq 0$$

në qoftë se p -vlera $< \alpha$.

- Në rastin e Tasty Sub numri i banorëve është është i rëndësishëm në nivel signifikance $\alpha = 0.05$, por jo në nivel $\alpha = 0.01$.
 - Rangimi i biznesit është i rëndësishëm në nivel $\alpha = 0.01$.
 - t_α , $t_{\alpha/2}$ dhe p -vlerat janë mbështetur në 7 shkallëlirë

15.6. Intervallet e besueshmërisë dhe ato të parashikimit

- Pika në vijën (sipërfaqen) e regresionit që i korrespondon një vlere të veçantë $x_{01}, x_{02}, \dots, x_{0k}$ të variablave të pavarura është

$$\hat{y} = b_0 + b_1 x_{01} + b_2 x_{02} + \dots + b_k x_{0k}$$

- Nuk ka shumë të ngjarë që kjo vlerë të jetë e barabartë me vlerën mesatare të y për këto vlera të x_1, x_2, \dots, x_k .
- Prandaj, duhet të vëjmë kujdes se sa larg mund të jetë vlera e parashikuar nga vlera aktuale.
- Bëjmë kërkime duke llogaritur një *interval besueshmërie* për vlerën mesatare të y dhe një *interval parashikimi* për një vlerë individuale të y .

Shembulli 15.5. Rasti i Tasty Sub

- 95%-intervali i besueshmërisë dhe 95%-intervali i parashikimit

Predicted values for: Të ardhurat vjetore, y (në \$1000)							
			<u>95% Confidence Interval</u>		<u>95% Prediction Interval</u>		
(në 1000)	znesit, x2	Predicted	lower	upper	lower	upper	Leverage
47.3	7	956.6057	921.0239	992.1876	862.8441	1,050.3673	0.168

15.8. Përdorimi i dummy-variablave për modelimin e variablave të pavarura kualitative

- Deri më tani kemi përfshirë vetëm të dhëna kuantitative në një model regresioni.
- Por, mund të përfshihen edhe të dhëna përshkruese kualitative.
 - Për shembull, mund të dëshirojmë të përfshijmë gjininë e të anketuarve
- Mund të modelojmë efektet e niveleve të ndryshme të një variableje kualitative duke shfrytëzuar të ashtuquajturat *dummy-variabla*.
 - Njihen gjithashtu edhe si variable indikatorë.

Shembull 15.6. Rasti i Electronics World

Shitorja	Amvisëri	Lokacioni	Shitje	DM	DD
1	161	Rrugë	157.27	0	0
2	99	Rrugë	93.28	0	0
3	135	Rrugë	136.81	0	0
4	120	Rrugë	123.79	0	0
5	164	Rrugë	153.51	0	0
6	221	Qendër tregtare (Mall)	241.74	1	0
7	179	Qendër tregtare	201.54	1	0
8	204	Qendër tregtare	206.71	1	0
9	214	Qendër tregtare	229.78	1	0
10	101	Qendër tregtare	135.22	1	0
11	231	Qendër e qytetit (Downtown)	224.71	0	1
12	206	Qendër e qytetit	195.29	0	1
13	248	Qendër e qytetit	242.16	0	1
14	107	Qendër e qytetit	115.21	0	1
15	205	Qendër e qytetit	197.82	0	1

Shembull 15.6. Rasti i Electronics World (Vazhdim)

Regression Analysis							
		R ²	0.437				
		Adjusted R ²	0.343	n	15		
		R	0.661	k	2		
		Std. Error	39.787	Dep. Var.	Sales		
ANOVA table							
	Source	SS	df	MS	F	p-value	
	Regression	14,716.2687	2	7,358.1343	4.65	.0320	
	Residual	18,995.8916	12	1,582.9910			
	Total	33,712.1603	14				
Regression output							
	variables	coefficients	std. error	t (df=12)	p-value	confidence interval	
	Intercept	132.9320	17.7932	7.471	7.52E-06	94.1639	171.7001
	DM	70.0660	25.1634	2.784	.0165	15.2397	124.8923
	DD	62.1060	25.1634	2.468	.0296	7.2797	116.9323

15.9. Përdorimi i variablave katrore dhe atyre të interaksionit

- Modeli kuadratik që lidh y me x është:

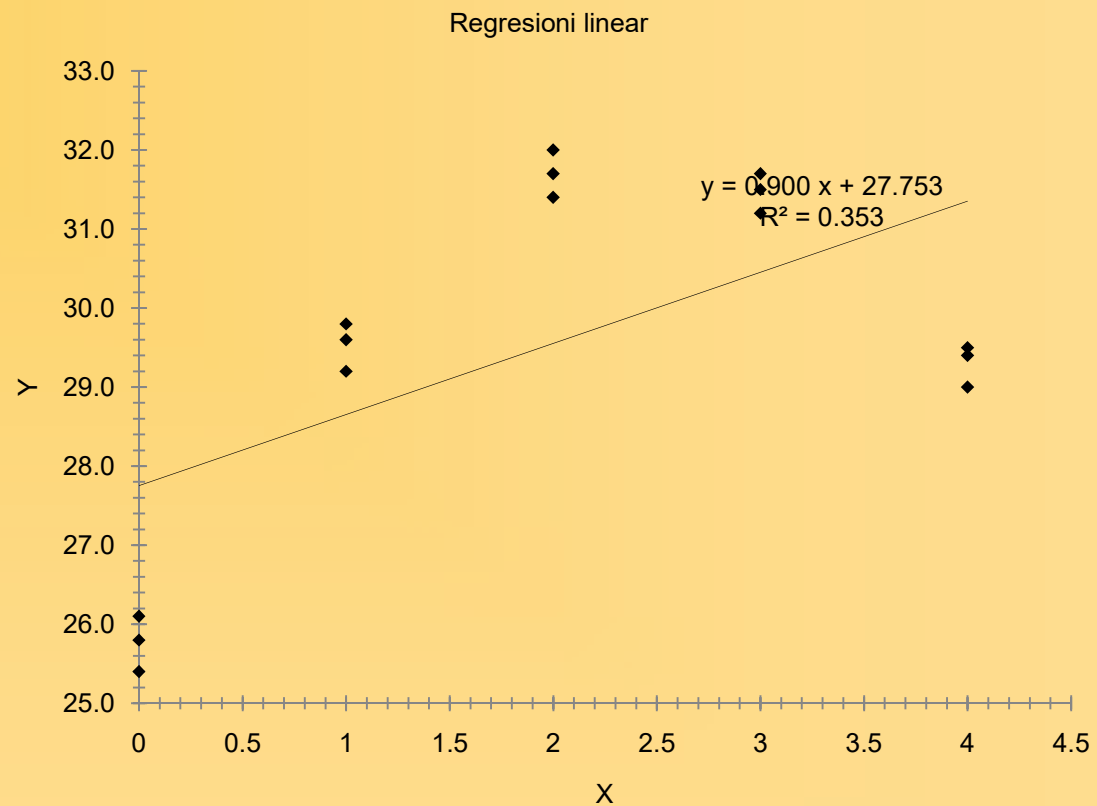
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

ku

- $\beta_0 + \beta_1 x + \beta_2 x^2$ është vlera mesatare e variablës së varur y
- Janë parametrat e regresionit që kanë të bëjnë me vlerën mesatare të y
- ε është term i gabimit që përshkruan se efektin në y të të gjithë faktorëve tjerë përveç x dhe x^2 .

Shembull 15.7. Rasti i aditivit të karburantit

Njësi	Mile
X	Y
0	25.8
0	26.1
0	25.4
1	29.6
1	29.2
1	29.8
2	32.0
2	31.4
2	31.7
3	31.7
3	31.5
3	31.2
4	29.4
4	29.0



Shembull 15.7. Rasti i aditivit të karburantit (Vazhdim)

Njësi		Mile
X	X^2	Y
0	0	25.8
0	0	26.1
0	0	25.4
1	1	29.6
1	1	29.2
1	1	29.8
2	4	32.0
2	4	31.4
2	4	31.7
3	9	31.7
3	9	31.5
3	9	31.2
4	16	29.4
4	16	29.0
4	16	29.5

Regression Analysis

	R ²	0.986		
Adjusted R ²	0.983	n	15	
R	0.993	k	2	
Std. Error	0.286	Dep. Var.	Y	

ANOVA table

Source	SS	df	MS	F	p-value
Regression	67.9152	2	33.9576	414.92	8.39E-12
Residual	0.9821	12	0.0818		
Total	68.8973	14			

Regression output

variables	coefficients	std. error	t (df=12)	p-value	95% lower
Intercept	25.7152	0.1554	165.431	1.60E-21	25.3766
X	4.9762	0.1841	27.025	4.05E-12	4.5750
X^2	-1.0190	0.0441	-23.085	2.60E-11	-1.1152

15.10. Ndërtimi i modeleve dhe efekti i multikolinearitetit

- Multikolineariteti është kushti kur variablat e pavarura janë të varura, në relacion ose në korelacion ndërmjet veti.
- Efektet:
 - Pengon aftësinë për përdorimin e p-vlerave për të vlerësuar rëndësinë relative të prediktorëve
 - Nuk e pengon aftësinë për të parashikuar variablën e varur
- Detektimi:
 - Matrica e paraqitjes grafike
 - Matrica e korelacionit
 - Faktorët e inflacionit të variablës (VIF)

15.11. Analiza e mbetjeve në regresionin e shumëfishtë

- Për një vlerë të observuar y_i mbetja është
$$e_i = y_i - \hat{y} = y_i - (b_0 + b_1x_{i1} + \dots + b_kx_{ik})$$
- Në qoftë se supozimet e regresionit janë të sakta, atëherë mbetjet duhet të duken si mostër e rastësishme nga një shpërndarje normale me mesatare 0 dhe devijim variancë σ^2 .

15.12. Regresioni logjistik

- Regresioni logjistik është shumë i ngjashëm me regresionin e katrorëve më të vegjël
 - Të dyja prodhojnë ekuacione parashikuese
- Regresionin logjistik e bën të ndryshëm variabla y .
 - Te regresioni i katrorëve më të vegjël variabla y është variabël kuantitative
 - Te regresioni logjistik, variabla y është dummy-variabël (0 ose 1)