

Statistika Kodutöö

14. Missugune suurus näitab, et regressioonimudel on parim?

Lühem vastus:

Veel kasutatakse: MSE, MAE, SSE, SAE, ristentroopia.

R² tähendab R-ruutu (korrelatsioonikordaja ruut, coefficient of determination).

Lühivastus konkreetse näite kohta: R². Parim funktsioon on suurima R²-ga.

Korrelatsioonikordaja väärtus võrdub R-ga (ruutjuur R²-st).

Täpsem vastus:

Ainult ühe suuruse järgi ei saa otsustada, milline regressioonimudel on parim, sest isegi täpse mudeli korral võivad rohkem R² suurenemisse panustavad sõltumatud muutujad olla statistiliselt mitteolulised.

Mudeli kvaliteedi oluline mõõt on see, kui suure osa sõltuva tunnuse dispersioonist saame seletada teise või teiste tunnuste (argumenttunnuste) kaudu.

Kuna mittelineaarsed mudelid saab taandada lineaarseteks, tehes vastavad teisendused

- a. If $y = ae^{bx}$, then $\ln y$ is linear with x .
- b. If $y = ax^b$, then $\log y$ is linear with $\log x$.
- c. If $y = L/(1 + e^{a+bx})$, then $\ln\left(\frac{L-y}{y}\right)$ is linear with x .
- d. If $y = \frac{1}{a+bx}$, then $\frac{1}{y}$ is linear with x .
- e. If $y = \frac{x}{a+bx}$, then $\frac{1}{y}$ is linear with $\frac{1}{x}$.
- f. If $y = 1 - e^{-x^b/a}$, then $\ln \ln\left(\frac{1}{1-y}\right)$ is linear with $\ln x$.

saame lineaarsed mudelid, mida võib väga hästi (ka mittelineaarseid võib, aga tõlgendus on keerulisem) võrrelda R² või taandatud R²-kordaja abil (kui on mitmemõõtmeline probleem)

Taandatud (adjusted) R² näitab seda, mis osa sõltuva muutuja dispersioonist seletavad tegelikult seda muutujat mõjutavad sõltumatud muutujad.

$$R_{adj}^2 = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

N on valimi suurus ja k on sõltumatute argumenttunnuste arv.

Kuna n-k-1 ei ole kunagi võrdne n-1-ga, siis taandatud R² ei ole kunagi võrdne R²-ga, v.a. siis kui ei ole ühtegi sõltumatut muutujat (peale vabaliikme).

Regressioonimudeli kvaliteedi analüüsimisel on olulisel kohal vead ehk prognoosijäägid (tegelik väärtus miinus prognoos). Jääkide keskmine on null ja mida väiksem on jääkide standardhälve, seda parem, seda vähem erinevad jäägid nullist ehk seda täpsem on prognoos. Enamik analüüsimeetodeid, nagu seletatud ka siin, eeldavad, et jäägid peavad olema tsentreeritud normaaljaotusega, sest enamike protsesside juures eeldatakse, et andmestik on genereeritud samade juhuslike faktorite poolt, mis on kõik täpselt sama jaotusega, ja sellisel juhul nende summa (ja samuti ka summa vigade jada) asümptootiline ($n \rightarrow \infty$) jaotus on normaaljaotus (tsentraalne piirteoreem). Kui analüüsi käigus selgub, et vead ei ole normaaljaotusega, siis võib olla, et andmepunkte genereerinud juhuslikud faktorid (suurused) on üksteisest sõltuvad (näiteks ajaliselt).

Eriti tähtis on vaadelda suuri jääke, mis võivad anda teada kas vigadest andmeis (saime mõttetult ebatüüpilise prognoosi) või ekstremaalsetest väärtustest (erindid või outliers).

18. Kirjuta tunnusele 'test' kirjeldus

18. Kirjuta tunnusele test kirjeldus	
test	
Mean	64,27586
Standard Error	1,517755
Median	66
Mode	73
Standard Deviation	14,15667
Sample Variance	200,4114
Kurtosis	0,089152
Skewness	-0,57339
Range	66
Minimum	26
Maximum	92
Sum	5592
Count	87
Confidence Level(95.0%)	3,017196

Tunnus test

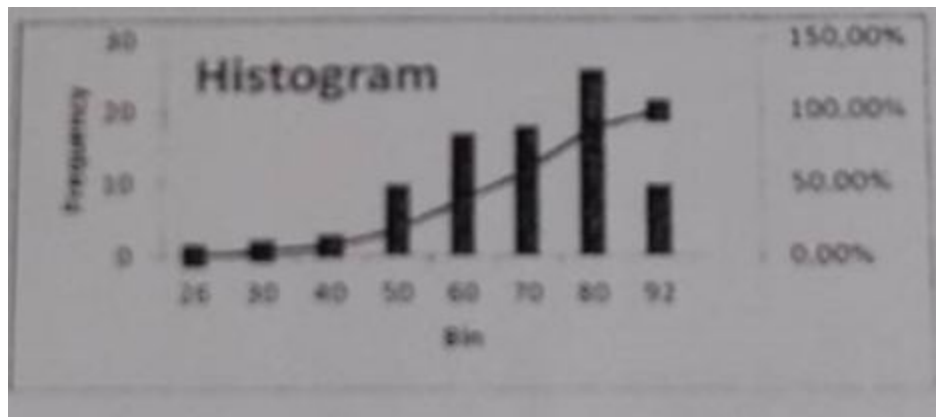
Tunnuse haare on 66 (miinimumi 26 ja maksimumi 92 vahel). Mood 73 näitab seda, et populaarseim (sagedaseim) tunnuse väärtus on 73. Väike (tunduvalt alla 3) positiivne järsakus

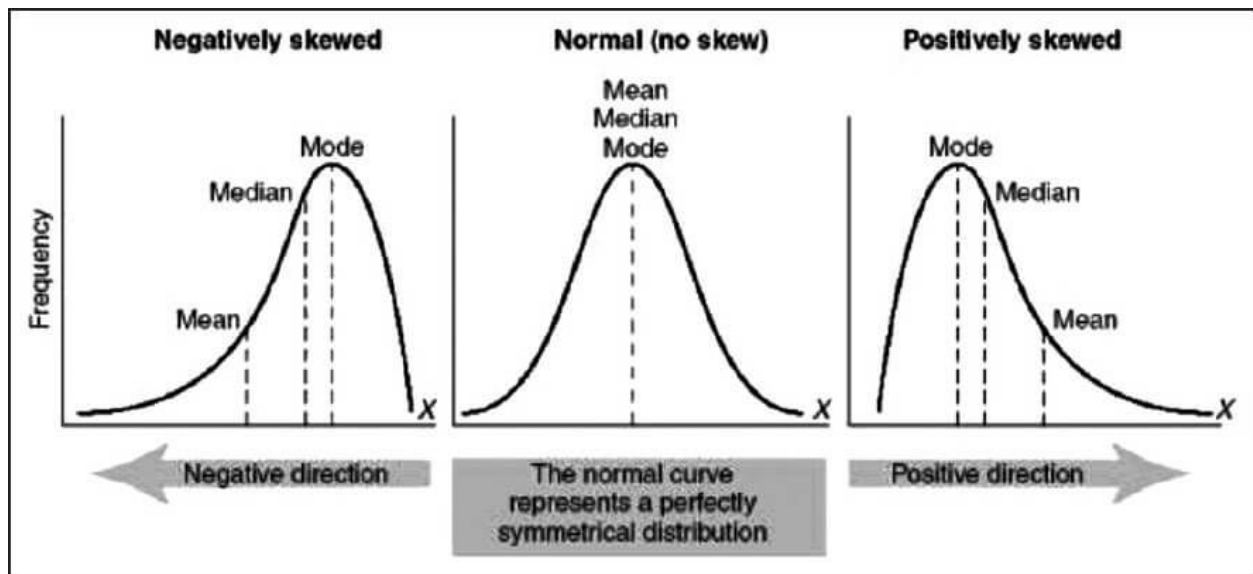
viitab platykurtic omadusele (ideaalne näite sellest on ühtlane jaotus), ehk antud suuruse tõenäosusjaotus annab tunduvalt vähem ja väiksemaid ekstreemsemaid väärtusi (hälbeid) kui normaaljaotus. Kuna keskvärtus on väiksem kui mediaan, on tegemist negatiivse asümmeetriaga.

Kui statistiku väärtus oleks 3.0, siis usaldusnivoole 95 % ($p=0.05$) vastav statistiku väärtus 3.0 tähendab seda, et tõenäosusega 95% üldkogumi parameetri tegelik väärtus ei erine statistiku väärtusest rohkem kui 3.0-kordse standardvea võrra.

Üksikmõõtmise standardhälve on ligikaudu 14.2 (ja selle ruut annab dispersiooni 200.4). Vaadeldava statistiku standardviga on ligikaudu 1.518 -- see annab teatava täpsushinnangu sellele statistikule.

Sum tähendab summat ehk siis tunnuse väärtuste summat. Seda saab kontrollida nii, et keskvärtus korda count (valimi suurus) peab andma selle summa (ja annab ka). Märgime, et antud andmete põhjal ei saa tunnuse tõenäosusjaotust üheselt määrata, aga nende andmetele vastava jaotuse näite saab leida. Kui 3.0 on 95% usaldusele vastav statistiku väärtus, siis kindlasti ei ole tegemist normaaljaotusega, sest oluliselt rohkem väärtusi on vähealt jaotuse ühes ääres (sel juhul parempoolses ääres). Kahjuks on kumulatiivne jaotus lehelt väljas.





21.

Reklaamiti hommikusöögi helbeid. Kas reklaamist oli kasu?

	<i>enne kampaaniat</i>	<i>peale kampaaniat</i>
Mean	4371,42857	4840
Variance	922380,952	846333,333
Observations	7	7
Pearson Correlation	0,77925224	
Hypothesized Mean Differen	0	
df	6	
t Stat	-1,98079768	
P(T<=t) one-tail	0,04745946	
t Critical one-tail	1,94318028	
P(T<=t) two-tail	0,09491893	
t Critical two-tail	2,44691185	

Missugused andmed: Andmed võivad olla näiteks erinevates piirkondades müüdnud hommikusöögipakkide müügist saadud käive ja keskmine on siis nende käivete keskmine. Statistikuks on käivete keskväärtuste erinevus enne ja pärast reklaami.
H0 Hüpotees: Käivete keskväärtused enne ja pärast reklaami on samad.

H1_{alt} hüpotees: Pärast reklaami müüdud pakkide arvu keskvärtus on suurem (ehk meid huvitab kas testi põhjal leitud t-statistik on väiksem kui vasakpoolne kriitiline t-statistik ja kas see erinevus on valitud usaldusnivool statistiliselt oluline). Olgu valitud usaldusnivoo $\alpha = 0.05$.

Testi tulemus:

Et hinnata, kas reklaamist oli kasu, uurime H₀-hüpoteesi kehtivust/mittekehtivust (mitte alternatiivse hüpoteesi kehtivust!)

Kuna suurema ja väiksema standardhälbe jagatis (rühmade põhjal) antud juhul on 1.04396, siis võib eeldada, et mõlema populatsiooni dispersioonid on võrdsed (sest see väärtus on alla 2).

Leiame, et andmete põhjal saadud t-statistiku väärtus on -1.98, aga vasakpoolne kriitiline t-statistiku väärtus on ligikaudu -1.94 -- seega antud andmete põhjal leitud t-statistik ekstremaalsem (väiksem) lubatud t-statistikust. Kuna p-väärtus 0.04 on väiksem kui valitud usaldusnivoo väärtus 0.05, siis on tõenäosus 1. Liiki viga teha väiksem lubatud tasemest ja võime H₀-hüpoteesi ühepoolsest ümber lükata.

Kokkuvõte: Reklaamist oli kasu. Lisaks kuna valim on väga väike (7), siis ühepoolse testi tulemuse valideerimiseks tuleks kindlasti kontrollida, kas keskmiste erinevus on normaaljaotusega või mitte, sest see on t-testi eelduseks.

22.

22. Uuritav tunnus on maksumus.		Kui suur on korrelatsioonikordaja: Missuguste tunnuste vahel on see leitud?				
Regression Statistics		Mudel on:				
Multiple R	0,9912	Kirjeldatuse aste on:				
R Square	0,9826					
Adjusted R Square	0,9814					
Standard Error	0,5254					
Observations	17					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	233,2365	233,2365	844,8276	1,33E-14	
Residual	15	4,141138	0,276076			
Total	16	237,3776				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-1,165	0,314365	-3,70473	0,002118	-1,83469	-0,49458
aastakaive	0,8632	0,029699	29,06592	1,33E-14	0,799914	0,926515
Mudeli interpretatsioon on:						

Korrelatsioonikordaja on võrdne multiple (mitmene) R-ga ehk 0.9912. See on leitud uuritava ja sõltumatu tunnuste lineaarkombinatsiooni vahel.

Mudel on: $\text{maksumus} = -1.165 (+0.67) + 0.8632(+0.063) \cdot \text{aastakäive}$
 Mudeli kirjeldatuse aste on 0.9826 ehk 98.26 % või 0.9814 ehk 98.14 %.

23.

23. Uuriti kahe firma patareide eluiga. Kas eluiga on erinev?		
t-Test: Two-Sample Assuming Equal Variances		
	A	B
Mean	131,5	99,3
Variance	6205,611	8907,122
Observations	10	10
Pooled Variance	7556,367	
Hypothesized Mean Difference	0	
df	18	
t Stat	0,828294	
P(T<=t) one-tail	0,209176	
t Critical one-tail	1,734064	
P(T<=t) two-tail	0,418352	
t Critical two-tail	2,100922	

Kas dispersioonid on võrdsed?
 Andmed on sõltuvad /sõltumatud
 Hüpoteesid on:

Uuringu tulemus on:

Kuna standardhälvete jagatis erineb vähem kui 2 korda (1.19805), siis võib ka dispersioonid lugeda võrdseteks (siis dispersioonid erinevad kindlasti vähem kui 4 korda). Alternatiivselt võib teha Welchi t-testi.

Andmete sõltumatuse küsimus: kuna valimid on erinevatest firmadest, siis on alust eeldada, on need valimid on omavahel sõltumatud (sest erinevates firmades toodetakse patareisid tõenäoliselt erinevat moodi), seda tuleks eraldi testida (kas $P(\text{firma1}) \cdot P(\text{firma2}) = P(\text{firma1}, \text{firma2})$). Andmed oleksid sõltuvad, kui võrreldaks sama firma patareide eluiga. Olgu valitud olulisusnivoo 0.05.

Hüpoteesid:

H0: Kahe firma patareide eluea keskväärtused on võrdsed.

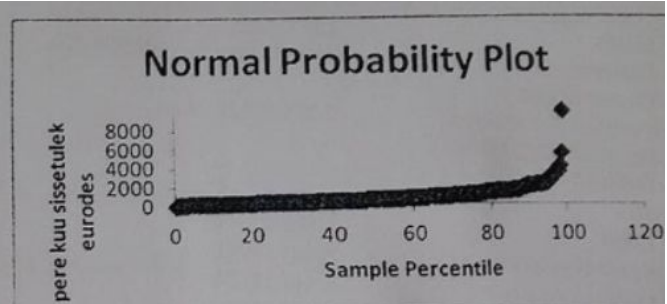
H1: Firma A patareide eluea keskväärtus on erinev.

Uuringu tulemus: Kuna andmete põhjal leitud t-statistik on oluliselt väiksem kahepoolsest t-statistikust (samuti p-väärtused alfa-tasemest suuremad), siis H0-hüpoteesi ümber lükata ei saa.

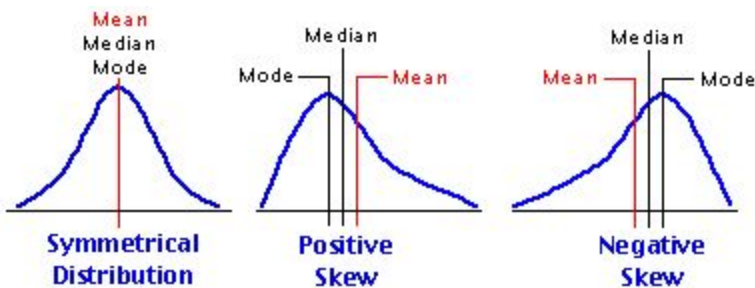
Uuringu tulemus: Usaldusnivool 0.05 ei saa ümber lükata hüpoteesi, et kahe firma patareide eluea keskväärtused on võrdsed. Jällegi tuleb ühepoolse testi tulemuste usaldusväärsuseks teha normaalsuse kontroll, sest valimi suurus 10 on oluliselt alla 30.

24.

24. Mida saame otsustada kõrvaloleva graafiku põhjal?



Graafiku põhjal saab järeldada, et antud tunnuse jaotuse standardhälve on võrdlemisi suur üksikute suurte väärtuste tõttu (maksimaalne väärtus ligi 9000). Keskvärtus on võrreldes maksimaalsega päris väike, ilmselt umbes 1500-kandis, mediaan on umbes 500 ja 1000 vahel ja miinimum 0-lähedane. See tähendab, et keskvärtus on tunduvalt mediaanist suurem -- see tähendab, et jaotuse asümmeetria on positiivne ja tegu on võrdlemisi raske sabaga jaotusega, st selle jaotuse saba võib olla raskem kui eksponentjaotusel.



25.

25. Mis tunnus sobib nendest valida uuritavaks tunnuseks?

	regioon	peremehe vanus	pere-liikmete arv	pere kuu sissetulek eurodes	peremehe sugu 1-meis	peremehe õpingu-aastaid
regioon	1					
peremehe vanus	-0,05718	1				
pere-liikmete arv	0,008646	-0,30603	1			
pere kuu sissetulek eurodes	-0,00562	-0,35224	0,256159	1		
peremehe sugu 1-meis	0,030467	-0,13799	0,413231	0,1565545	1	
peremehe õpingu-aastaid	0,152151	-0,32419	0,051822	0,3290712	0,072386	1

Mis tunnused sobivad valida argumenttunnusteks?

Mille põhjal toimub valik? Kas on mingeid kitsendusi?

Tunnuste valimine sõltub sellest, kas on ühene või mitmene regressioon.

Mitmese regressiooni korral on heade uuritava ja argumenttunnuste valimine on keerulisem, sest korrelatsioon ei tähenda kausaalsust. St ühe muutuja väärtused võivad olla mõjutatud teiste muutujate väärtustest (läbi interaktsioonide) ja seda on korrelatsioonimaatriksist veidi tülikas välja lugeda -- kasulik oleks ehitada graafiline mudel. Argumenttunnuse valimisel tasub vältida selliseid, mis tugevalt omavahel korreleeruvad (multikollineaarsus). Vaadates korrelatsioonimaatriksit on selge, et regioon ise kõrgelt teiste muutujatega ei korreleeru. Kuna peremehe vanus tundub korreleeruvat võrdlemisi tugevalt enamike muutujatega v.a regioon, siis võiks uuritavaks tunnuseks valida peremehe vanuse, sest seda tunnust on argumenttunnusena potentsiaalselt lihtsam asendada (sest on suurem tõenäosus, et see muutuja on kollineaarne teistega). Mitmese regressiooni korral kahe muutuja vahelise korrelatsiooni teadmist ei piisa, sest mitmesel regressioonil võib nõrgalt omavahel korreleerunud argumenttunnuste lineaarkombinatsioon korreleeruda uuritava tunnusega tugevamini kui lineaarkombinatsioon tunnustest, mis eraldi võttes uuritava tunnusega kõrgelt korreleeruvad -- selle seletuseks võib olla Simpsoni paradoks.

Kui eesmärgiks on nõ tavaline regressioon, siis uuritava tunnuse valimisel tuleks vältida kategoorilisi tunnuseid (siis on tegu logistilise regressiooniga ja väljundit tõlgendatakse aktivatsioonitõenäosusena, mitte väljundi enda väärtusena, sest kategooriline muutuja ei saa omada mistahes väärtusi ta enda väärtuste vahel). Peremehe sugu, pereliikmete arv ega regioon sel juhul uuritavaks tunnuseks ei sobi, küll sobiksid need hästi indikaatortunnustena argumenttunnusteks. Selles ülesandes võib argumenttunnuseks võtta mistahes tunnuse.