Otázky k tématu 6 – regrese[1]

1. Vyberte správnou odpověď

1.1 Který termín patří mezi ostatní nejméně?
a) percentil
b) korelace
c) regrese
d) predikce

1.2 O lineárním vztahu mezi dvěma proměnnými….
a) … vypovídá korelace více než regresní rovnice
b) … vypovídá korelace méně než regresní rovnice
c) … vypovídají korelace i regresní rovnice stejně
d) … podávají regresní rovnice a korelace různé informace

1.3 K čemu slouží linearizační transformace? V jakém případě o ní uvažujeme?

2. Následující otázky se týkají obecných aspektů regresní analýzy

2.1 Jaké jsou dvě důležitá použití regresní analýzy?

2.2 Když použijete směrodatnou odchylku reziduálních hodnot jako ukazatel chyby predikce, proč je důležitý předpoklad homoskedascity?

2.3 Co znamená, že regresní přímka je založená na metodě nejmenších čtverců?

2.4 Proč nemůže být $\sum(Y - Y')$ použita jako ukazatel chyby odhadu?

2.5 Popište, jakým způsobem se *r* vztahuje k predikci/odhadu chyby.

2.6 Platí, že s rostoucí hodnotou *r* roste i směrodatná odchylka reziduí (reziduální rozptyl)?

2.7 Proč musí být výzkumník opatrný, použije-li rovnici predikce u jedince, který má odlišné charakteristiky od lidí, kteří byli zahrnuti v originálním vzorku, z kterého byla rovnice regrese odvozena?

3. Určete správné odpovědi

3.1 Jestliže $r = 0$, čemu musí být rovno b?

3.2 Je-li $r = 1$, znamená to, že ve všech párech hodnot korelovaných proměnných jsou obě hodnoty stejné?

3.3 Pokud $r = 0{,}5$ a $z_X = 2{,}0$, kolik je odhad hodnoty Y, tj. z_Y' ?

3.4 Jakému percentilu odpovídá hodnota z_Y' z předchozího příkladu? Předpokládejte, že rozložení proměnných je normální.

3.5 Pokud $r = -0{,}6$ a $z_X = -1{,}5$, kolik je odhad hodnoty Y, tj. z_Y' ?

3.6 Když predikujeme Y z X a $s_X = s_Y = 15$, platí že korelační koeficient (r) je roven regresnímu koeficientu (b)?

3.7 Pokud $s_Y = 10$ a $r = 0{,}6$, jaká je směrodatná odchylka reziduálních hodnot (s_e)?

---

1 Pro označení hodnot predikované proměnné používám níže apostrof (Y')

3.8 Pokud r_{XY} = 0,5, a výkon osoby u proměnné X odpovídá P_2, odhadněte na úrovni jakého percentilu bude výkon téže osoby u proměnné Y.

a) P_{50}
b) P_{75}
c) P_{16}
d) P_2

3.9 Pokud s_Y = 10 a r = 0,6, jaký je rozptyl predikovaných hodnot Y'

3.10 Pokud s_Y = 10 a r = 0,6, jaká je směrodatná odchylka reziduálních hodnot (s_e, resp. s_{res})?

4. Předpokládáme-li bivariační normální rozložení

4.1 Pokud směrodatná odchylka reziduálních hodnot s_e = 8, kolik procent skutečných hodnot závislé proměnné se od předpovězené hodnoty lišit o méně než 8 bodů?

4.2 A kolik procent skutečných hodnot bude o více než 8 bodů vyšší než předpovězené hodnoty?

4.3 Bude podíl předpovědí, které podhodnotí skutečnou hodnotu o více než 8 bodů, stejný jako v předchozí otázce?

5. Korelace mezi IQ skóry kteréhokoli z rodičů a jejich dětí je přibližně 0,5. Také víme, že průměr IQ rodičů i dětí je stejný (100) a směrodatné odchylky v obou populacích jsou také stejné (15).

5.1 Odhadněte průměrné IQ dětí matek s IQ = 130

5.2 Odhadněte průměrné IQ dětí otců s IQ = 90

5.3 Odhadněte průměrné IQ dětí matek s IQ = 100.

5.4 Průměr IQ obou rodičů koreluje s IQ jejich dítěte přibližně 0,6. Jaká bude směrodatná odchylka chyb odhadu (s_e, s_{res}) při predikci IQ dětí?

5.5 Pokud s_{res} = 12, v jakém procentu případů se budou skutečné IQ skóry lišit od předpovězených o více než 12 bodů.

6. Ve studiu tolerance k bolesti se výzkumník zajímá o predikci/odhad doby, během které jsou subjekty schopni udržet ruce ve velmi ledové vodě. Z dřívějšího výzkumu ví, že příjem vitamínu E během minulých 12 hodin koreluje s tolerancí bolestivého stimulu, přinejmenším, pokud je jím zmrzlá voda. Následující tabulka ukazuje páry skórů pro studovaný vzorek:

Vitamin E: (X)	*Tolerance Times (in seconds): (Y)*
5	23
9	32
22	65
12	40
16	42

6.1 Jaká je směrnice b?

6.2 Jaký je průsečík a?

6.3 Jaké je směrodatná odchylka chyb odhadu s_e ?

6.4 Jaký čas tolerance byste predikovali někomu, kdo si vzal ráno při studii 16 jednotek vitamínu E?

7. Sociolog se zajímá o predikci ročního platu (Y) založenou na dřívější úrovni vzdělání (X). Úroveň vzdělání je definována jako počet let školní docházky. Následující data byla získána od 6 subjektů:

Education: (X)	Income X 1000: (Y)
10	15
14	29
9	14
14	37
12	20
13	23

7.1 Jaká je směrnice b?

7.2 Jaký je průsečík *a*?

7.3 Jaké je směrodatná odchylka chyb odhadu s_e ?

7.4 Jaký příjem byste odhadovali pro někoho s desetiletou docházkou?

8. Vstupní komise potřebuje odhadnout, zda určitý student bude schopný získat potřebné známky během prvního roku na vysoké škole. Aby prošel prvním rokem, musí dosáhnout 3,00 GPA (průměrný prospěch, vyšší čísla lepší) v prvním semestru na univerzitě. Vstupní komise má data z dřívějších let, a to GPA ze střední školy a GPA z prvního roku na univerzitě. Data jsou následující:

Undergraduate GPA: (X)	Graduate School GPA: (Y)
3.50	3.33
3.98	3.63
3.10	3.40
2.90	3.41
3.40	3.40

8.1 Jaké je *b*?

8.2 Můžete přijmout studenta se středoškolským GPA = 3,00?

8.3 Mezi jakými hodnotami výsledků GPA v prvním roce na univerzitě by se mělo ocitnout 68% studentů, kteří vstupovali do studia se středoškolským GPA = 3,67?

9. Následující příklady vycházejí z údajů zjištěných reálnými výzkumy. Pro jejich výpočet použijte statistický software.

9.1 Baron, Logan a Kao (1990) studovali vztah mezi nepohodlím pacientů, které bylo posuzováno studenty stomatologie a samotnými pacienty. Nepohodlí bylo definováno jako kombinace úzkosti, bolesti a distresu (*nižší hodnoty/čísla indikují nižší nepohodlí*). Jednotlivé hodnoty byly získány za dvou podmínek: během vrtání a během umístění pryžové hráze kolem zubu. (Pryžová hráz je gumová pochva okolo zubu, která je připojena na kovový rámec, izoluje zub a tím brání tomu, aby byly úlomky spolknuty. Umístění této hráze vyžaduje zvýšenou pozornost ze strany zubaře než jednoduchá výplň zubu). Korelace mezi nepohodlím posuzovaným studenty stomatologie a samotnými pacienty byly spočítány za obou podmínek. Korelace mezi hodnoceními během vrtání byly významné: $r(41) = +0{,}52$, $p < 0{,}05$. Korelace během umístění hráze mezi hodnocením studenty stomatologie a pacienty byly také významné, ale o hodně menší: $r(41) = +0{,}21$, $p < 0{,}05$. Autoři spekulují, že schopnost postihnout distres pacientů závisí na tom, co dělají… „a že být senzitivní vůči stresu vyžaduje značnou kapacitu pozornosti (s. 151).“ Vypočítejte rovnici regrese a s_e pro obě podmínky. Jestliže během vrtání student stomatologie ohodnotí pacientovo nepohodlí 7, jak byste odhadli/predikovali hodnocení nepohodlí samotným pacientem? Odpovězte na stejné otázky i pro druhou podmínku a určete směrodatnou odchylku chyb odhadu.

Discomfort Ratings During Drilling

Dental Students	*Patients*
8	6
6	9
3	1
1	4
5	5
4	6
8	8
7	6
9	6
2	3
1	1
6	8
4	6
3	3
9	7
7	8
6	9
2	8
5	7
6	6
3	2
1	1
5	7
6	9
8	8
9	6

Discomfort Ratings During Rubber Dam

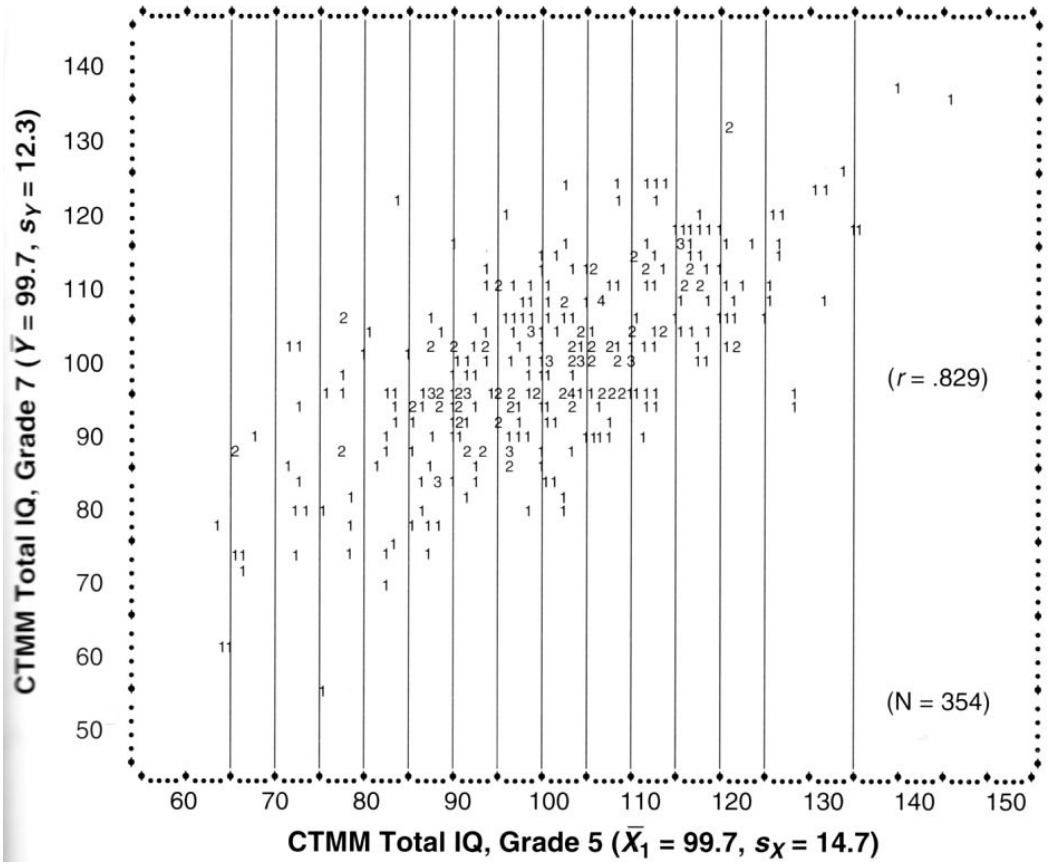
Dental Students	*Patients*
8	6
6	9
3	1
1	4
5	5
4	6
8	8
7	6
9	6
2	3
1	1
6	8
4	6
3	3
9	7
7	8
6	9
2	8
5	7
6	6
3	2
1	1
5	7
6	9
8	1
9	4

9.2. Carrie (1981) zkoumala vztah mezi podáváním zpráv o symptomech během těhotenství a během menstruace a asociací těchto zpráv s obecnou tendencí podávat zprávy/vypovídat o psychologických a fyzikálních symptomech. Kromě jiného zjistila významnou korelaci mezi počtem symptomů zažívaných během menstruace a počtem symptomů, o kterých podávaly zprávy během těhotenství. Pro ženu, která vypovídá o 54 menstruačních symptomech, určete, o kolika symptomech bude vypovídat během těhotenství? Identifikujte dvě hodnoty těhotenských symptomů, mezi kterými se nachází 68 procent žen, které vypovídají o 54 menstruačních symptomech.

Hypothetical Questionnaire Scores

Last Menstruation Symptoms	*Last Pregnancy Symptoms*
93	87
75	64
34	78
23	55
76	43
34	45
21	20
34	54
60	60
45	82
67	67
50	48
89	72
61	68
56	45
82	75
45	34
53	55
71	50
59	90
90	56
43	62
49	32

10. Následující obrázek je vygenerovaný scatter IQ skórů 354 dětí testovaných v 5. (X) a 7. (Y) třídě.



10.1 Jaký je nejvyšší a nejnižší skór v 5. třídě?

10.2 Jaký je nejvyšší a nejnižší skór v 7. třídě?

10.3 Vypadá regrese lineárně?

10.4 Splňuje podle scatteru závislá proměnná podmínku homoscedascity?

10.5 Spočítejte regresní koeficient $b_{Y,X}$.

10.6 Spočítejte průsečík *a*.

10.7 Vytvořte regresní rovnici s dosazenými koeficienty.

10.8 Tomáš získal v 5. třídě skór IQ 140. Předpovězte jeho skór IQ v 7. třídě.

10.9 David získal v 5. třídě skór IQ 70. Předpovězte jeho skór IQ v 7. třídě.

10.10 Zakreslete do obrázku regresní přímku.

10.11 Spočítejte směrodatnou odchylku chyb odhadu s_e (s_{res}).

10.12 Jaká část predikovaných skórů se bude lišit od skutečných hodnot o méně než 7 bodů?

10.13 S přibližně dvoutřetinovou pravděpodobností leží Tomášovo IQ v sedmé třídě mezi ____ a ____ a Davidovo mezi ____ a ____.

11. Konkrétní test inteligence koreluje s testem čtení 0,82. Víte-li, že m_{IQ}=100, s_{IQ}=15, $m_{Č}$=8 a $s_{Č}$=2 a předpokládáme, že obě proměnné jsou normálně rozložené.

11.1 Určete regresní rovnici na predikci čtení ze skórů IQ.

11.2 Jaký je průměrný výkon ve čtení při IQ=100?

11.3 Jaký je průměrný výkon ve čtení při IQ=90?

11.4 Porovnejte percentilové ekvivalenty skórů X a Y' z předchozí podúlohy.

11.5 Kolik % dětí s IQ=90 bude mít nadprůměrný skór ve čtení? [nápověda: cesta vede přes reziduální(chybový) rozptyl]

12. V odborné studii zabývající se chatováním na internetu jsme se dočetli, že vztah mezi chatováním a mírou depresivity se dá vyjádřit regresní rovnicí $\hat{y} = x$, kde *y* reprezentuje míru chatování (počet hodin strávených týdně chatováním) a *x* reprezentuje skóre depresivity. Studie rovněž uvádí základní popisné statistiky obou proměnných:

	N	m	s
chatování	2380	0,6	2,0
depresivita	2380	1,6	0,8

12.1 Jak velký podíl rozptylu chatování lze vysvětlit depresivitou?

12.2 Kdybychom chtěli naopak predikovat depresivitu z míry chatování, jak by vypadala regresní rovnice?

12.3 Jaký skór depresivity bychom predikovali člověku, který strání týdně 10 hodin chatováním?

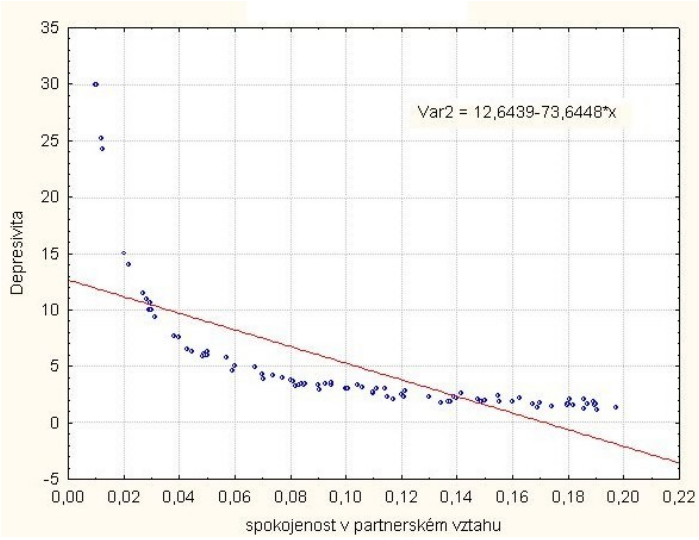
13. Vztah mezi průměrnou známkou u bakalářských státnic (BC na škále 1(=A) – 5(=E)) a výsledkem u písemné části přijímací zkoušky do magisterského navazujícího studia (MG na škále od 0 do 100) je popsán lineárně-regresní rovnicí *MG*' = 72 – 4*BC*.
Popisné statistiky:

- Průměrná známka u bakalářských státnic má přibližně normální rozložení s *M*=3 a *SD*=1,5
- Výsledky písemné části přijímací zkoušky mají přibližně normální rozložení s *M*=60 a *SD*=10.

13.1 Jaký výsledek budeme očekávat od uchazeče, který dosáhl u bakalářských státnic průměru 1?

13.2 Jaká je pravděpodobnost, že tento uchazeč s vynikajícím výsledkem u státnic nedosáhne u písemné části přijímací zkoušky požadované hranice 60 bodů (a nebude tedy přijat)? Řešení vede přes chybový rozptyl.

14.Následující graf zobrazuje vztah dvou proměnných, spokojenosti v partnerském vztahu a depresivity.



14.1 Jaký termín používáme pro tento typ grafu?

14.2 Korelace spokojenosti a depresivity je zde −0,7. Jaká je velikost účinku spokojenosti na depresivitu?

14.3 Převeďte velikost účinku na Cohenovo *d*.

14.4 Je-li směrodatná odchylka depresivity 8, jaký je rozptyl chyb odhadu depresivity prostřednictvím regresní rovnice uvedené v obrázku?

14.5 Popisuje lineární regrese dobře vztah těchto dvou proměnných? Proč?

14.6 Lze vztah, zde popsaný regresní rovnicí *depresivita* = 12,6 – 73,6*spokojenost*, slovy popsat také tak, že s každým nárůstem spokojenosti o jednu desetinu bodu očekáváme pokles depresivity o cca 7 bodů?

15. Ve studii tolerance k hluku se výzkumník zajímá o predikci doby, po kterou jsou zkoumané osoby schopny vydržet velmi hlasitý nepříjemný zvuk. Ví, že sluch je velmi adaptibilní, takže hypotetizuje, že čím více nahlas poslouchá adolescent svůj osobní přehrávač, tím delší dobu snese onen nepříjemný zvuk. Následující tabulka ukazuje páry skórů pro studovaný vzorek:

hlasitost přehrávače [*% maximální hlasitosti*]	**délka strpění nepříjemného zvuku [*s*]**
25	5
31	9
55	20
42	13
47	18
m = 40 *s = 12*	*m = 13* *s = 6*

15.1 Vytvořte bodový graf s hlasitostí jako prediktorem).

15.2 Pearsonův korelační koeficient mezi hlasitostí a délkou strpění nepříjemného zvuku činí 0,98. Jaké budou hodnoty Spearmanova a Kendallova korelačního koeficientu?

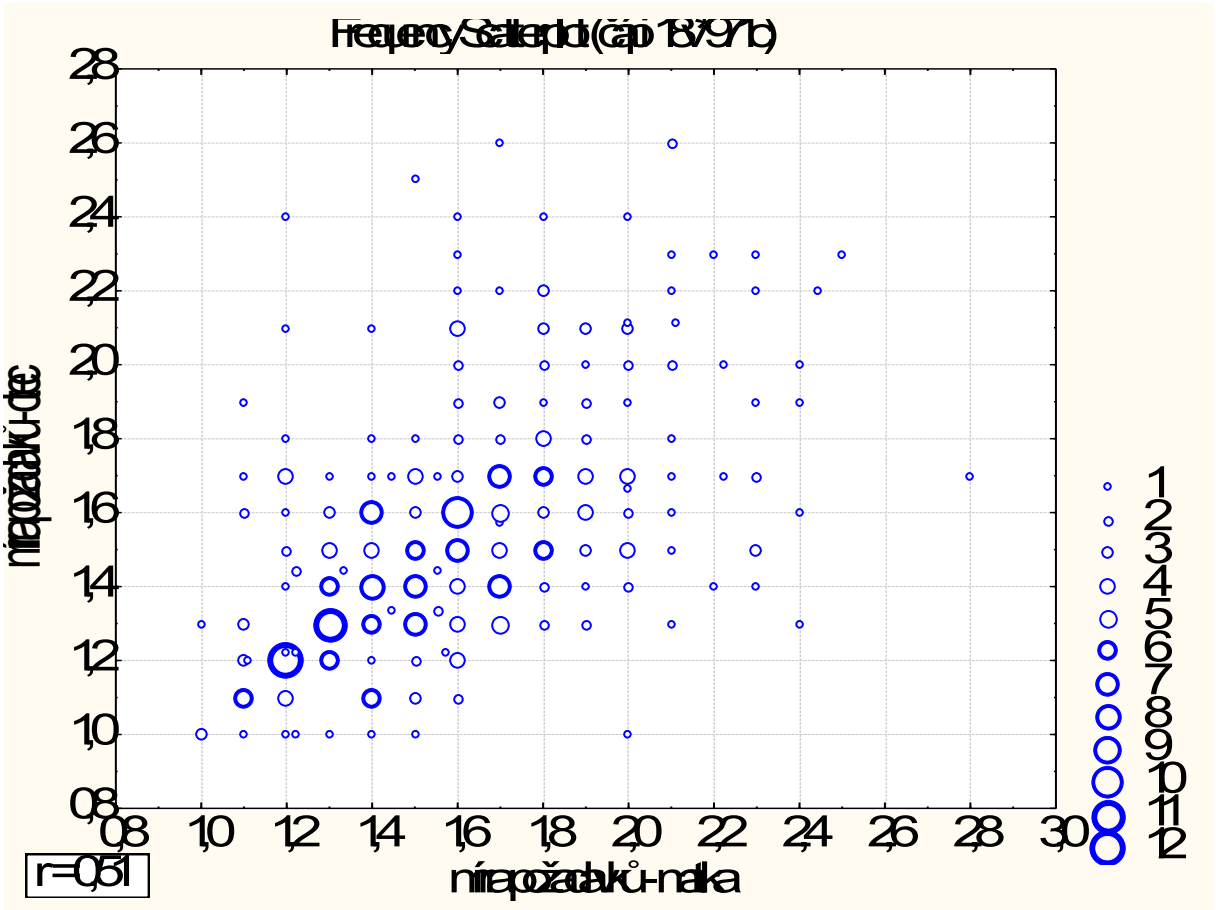
15.3 Vytvořte lineárně-regresní rovnici pro predikci délky strpění nepříjemného zvuku z preferované hlasitosti přehrávače.

15.4 Jaká je směrodatná odchylka chyb odhadu s_e (s_{res})?

15.5 Jakou délku strpění nepříjemného zvuku byste predikovali někomu, kdo poslouchá přehrávač na 60% maximální hlasitosti?

16. Následující graf popisuje souvislost proměnných „míra požadavků vůči dítěti“ ze strany matky a ze strany otce.

míra požadavků – matka: m_X = 1,6 s_X = 0,33
míra požadavků – otec: m_Y = 1,5 s_Y = 0,31 r_{XY} = 0,5



16.1 Uveďte předpoklady smysluplného použití lineární regrese. Které z nich můžeme považovat za splněné?

16.2 Vypočtěte parametry regresní funkce předpovídající míru požadavků otce z míry požadavků matky a regresní funkci zapište.

16.3 Proložte grafem vypočtenou regresní přímku.

16.4 Jaká část rozptylu proměnné Y je regresí vysvětlena?

16.5 Jaká je směrodatná odchylka odhadů míry požadavků u otce?

16.6 Jaký je průměr a směrodatná odchylka chyb odhadu?

17. Zajímá nás vztah proměnných „záporný emoční vztah“ (*ZEV*) a „míra požadavků“ (*MP*) ze strany matky vůči dítěti. Lineárně regresní funkce předpovídající míru požadavků ze záporného emočního vztahu má podobu *MP'* = 0,68 + 0,71*ZEV* a vysvětluje 40% rozptylu míry požadavků.

záporný emoční vztah:	m_X = 1,3	s_X = 0,28
míra požadavků:	m_Y = 1,6	s_Y = 0,31

17.1 Jaká je korelace mezi *ZEV* a *MP*?

17.2 Jaký je průměr a směrodatná odchylka reziduálních skórů (tj. chyb odhadu)?

17.3 Jaká je pravděpodobnost, že u náhodného dítěte nebude velikost chyby odhadu míry požadavků pomocí uvedeného regresního vztahu vyšší než +-0,24 bodu na škále míry požadavků?

18. Studentka Eva psala diplomku o autobiografické paměti manželů. Chtěla se dozvědět, co je pravdy na tom, že ženy si pamatují více z historie vztahu než jejich manželé. Ptala se jich tedy na to, kdy a kde se seznámili, kdy a kde si dali první pusu, kdy a kde měli poprvé sex apod. Celkem se každého zeptala na 12 takových událostí a zaznamenala si, u kolika z nich si pamatovali datum a u kolika místo. Také se ptala na délku současného vztahu v letech. Celkově to dalo dost práce, a tak máme data zatím ze 6 rodin:

Janů

Délka vztahu = 4	
Manželka	Manžel
data=9	data=3
místa=12	místa=6

Petrů

Délka vztahu = 5	
Manželka	Manžel
data=7	data=2
místa=9	místa=9

Vojtů

Délka vztahu = 8	
Manželka	Manžel
data=7	data=8
místa=3	místa=12

Jirků

Délka vztahu = 11	
Manželka	Manžel
data=6	data=7
místa=6	místa=11

Mirků

Délka vztahu = 14	
Manželka	Manžel
data=8	data=6
místa=4	místa=6

Nechoďdomů

Délka vztahu = 18	
Manželka	Manžel
data=11	data=4
místa=8	místa=10

Zajímá nás vztah mezi manželčiným „počtem zapamatovaných míst“ a „délkou vztahu“. Pokud tam nějaký vztah je, chtěli bychom predikovat délku vztahu z počtu míst, které si manželka pamatuje.

18.1 Nejprve vztah graficky zobrazte a slovně popište.
18. 2 Pomocí lineární regrese se pokuste predikovat délku vztahu v rodině, kde si manželka pamatuje jen 2 místa událostí.
18.3 Jaký je průměr a rozptyl reziduálních hodnot (reziduí)?
18.4 Jaká je pravděpodobnost, že se budeme v odhadu mýlit o více než 5 let?

19. Ve studii o kreativitě se zjišťovalo, nakolik ji ovlivňuje vizuální paměť. Administrovali jsme test vizuální paměti, jehož skór udává, kolik z předložených 20 objektů si je účastník schopen po 10 minutách vybavit. Kreativita měřená na intervalové škále měla *M*=40 a *SD*=10 a vizuální paměť měla *M*=13 a *SD*=6. Korelace mezi kreativitou a vizuální pamětí je 0,6.
19.1 Pomocí lineární regrese odhadněte kreativitu člověka, který má skór vizuální paměti 8.
19.2 Jaká je přibližně pravděpodobnost, že se ve svém odhadu budeme mýlit o více než 4 body?
19.3 Jakým způsobem zjišťujeme splnění předpokladu homoskedascity?
19.4 Doplňte větu interpretující regresní model predikující kreativitu z vizuální paměti:
S každým nárůstem …… o jednotku naroste odhad …… o …… bodů.

20. Jaké dva základní postupy můžeme zvolit, abychom mohli model lineární regrese použít i v případě zjevné nelinearity vztahu?

21. Aplikujeme-li regresní model i pro hodnoty prediktorů, které jsou mimo variační rozpětí dat původně použitých pro stanovení parametrů regresního modelu, mluvíme o

a) transformaci
b) generalizaci
c) interpolaci
d) extrapolaci
e) transpozici
f) interferenci
g) inferenci
h) externalizaci
i) aplikaci

22. Regrese tolerance na věk má podobu $Y=0{,}22X+15{,}6$. To znamená, že když ..... stoupne o 10 jednotek, budeme člověku odhadovat o ..... jednotky vyšší ..... (doplňte)

23. Jaké jsou následky nesplnění jednoho z jeho předpokladů lineární regrese – homoskedascity?

24. Co jsou to *sumy čtverců*?

25. Homoscedascita je predpokladem lineárnej regrese. Z čoho usuzujeme, že je tento predpoklad porušen?

26. Regresná rovnica znie: Y' = 3X – 2. Aké bude predikované skóre Y ak X je rovné 4?

27. Aké je najpoužívanejšie kritérium pre určenie regresnej priamky?
a) priamka prechádza cez väčšinu bodov
b) nad priamkou sa nachádza rovnaký počet bodov ako pod priamkou
c) priamka minimalizuje veľkosť reziduálnych hodnôt pre hodnoty x_i a y_i, ktorými priamku prekladáme

28. Priemer premennej X je 3 a priemer premennej Y je 7. Regresná priamka, ktoré predikuje Y bude prechádzať bodom (3;7). Je tento výrok pravdivý?

29. Bod, v ktorom pretína regresná priamka vertikálnu os je:
a) vždy priemerným skóre prediktoru
b) niekedy negatívny, niekedy pozitívny
c) vždy nula

30. Sklon regresnej priamky:
a) je rovnaký ako korelačný koeficient
b) môže byť pozitívny alebo negatívny
c) obe možnosti

31.Máme k dispozícii dáta v nasledujúcej tabuľke:

extraverzia	12	14	16	17	19	23	25	27
počet cigariet za deň	0	0	10	12	11	18	30	32

Vypočítajte priesečník a smernicu pre fajčenie cigariet ako závislej premennej (Y).

)