M6120 – 9. CVIČENÍ : **M6120cv09** (*Multikolinearita a korelovaná pozorování*)

A. Multikolinearita a její zdroje.

Multikolinearitou se rozumí vzájemná lineární závislost vysvětlujících proměnných. Přesnou multikolinearitou se rozumí případ, kdy jednotlivé sloupce $\mathbf{x}_j, j = 1, \ldots, p$ matice plánu $\mathbf{X}^* = (\mathbf{1}_n, \mathbf{X})$ jsou lineárně závislé, takže pro aspoň jednu nenulovou konstantu c_j platí

$$c_1\mathbf{x}_1 + ... + c_k\mathbf{x}_p = \mathbf{O}_n.$$

V praxi bychom se s tímto případem neměli setkávat, neboť při rozumně sestaveném regresním modelu využijeme lineární kombinaci a zmenšíme počet vysvětlujících proměnných. Podobně nereálný je v praxi případ ortogonálních vysvětlujících proměnných, kdy matice $\mathbf{X}$ je ortogonální a platí, že

$$\mathbf{X}'\mathbf{X} = \mathbf{I}_p, \quad k = p + 1, \quad \text{matice plánu} \quad \mathbf{X}^* = (\mathbf{1}_n, \mathbf{X}).$$

Multikolinearitou se rozumí případ, kdy přibližně platí rovnice vyjadřující lineární kombinaci vysvětlujících proměnných. V případě silné multikolinearity je determinant informační matice $\mathbf{X}'\mathbf{X}$ blízký nule, nejmenší vlastní číslo je rovněž blízké nule a matice $\mathbf{X}'\mathbf{X}$ je "skoro singulární".

Důvody multikolinearity mohou být různé:

- Multikolinearitu způsobuje regresní rovnice obsahující **nadbytečné vysvětlující proměnné**. Statistickými technikami můžeme přebytečné proměnné identifikovat a vyloučit z regresní rovnice.
- Multikolinearitu jen ztěží odstraníme v úlohách, kdy vzájemná spřaženost hodnot vysvětlujících proměnných je způsobena neuvažovanými veličinami nebo **formou statistického zjišťování**. Jde-li např. o údaje z časových řad, je podobný vývoj sledovaných veličin dostatečným důvodem vzniku multikolinearity. Vzhledem k tomu, že multikolinearitu hodnotíme výhradně na základě určitého souboru pozorování, stačí nesprávný výběr kombinací hodnot vysvětlujících proměnných, nereprezentujících obor možných hodnot, k existenci významné multikolinearity.
- Závažným důvodem multikolinearity je **skutečný vztah vysvětlujících proměnných** v rámci sledovaného jevu, procesu nebo systému. V tomto případě je třeba využít všechny informace nevýběrového charakteru k zlepšení kvality regresních odhadů.

B. Důsledky multikolinearity.

V případě **přesné multikolinearity** je matice $\mathbf{X}'\mathbf{X}$ singulární a běžnou inverzí nepořídíme odhad neznámých parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců.

Pro **přibližnou (silnou) multikolinearitu** jsme sice schopni matici $\mathbf{X}'\mathbf{X}$ invertovat, ale kvalita pořízených odhadů je poměrně nízká.

Snížení kvality se projeví

- v kovarianční matici $var(\hat{\boldsymbol{\beta}})$ a
- v přesnosti prováděných výpočtů.

Diagonální prvky matice

$$(\mathbf{X}'\mathbf{X})^{-1},$$

tj.

$$a_{jj} = diag(\mathbf{X}'\mathbf{X})^{-1}$$

(označované v literatuře jako **VIF - *variance inflarion factors***) úzce souvisí s **vícenásobnými korelačními koeficienty**, vyjadřující vztah j–té vysvětlující proměnné a lineární funkce ostatních vysvětlujících proměnných. Lze je zapsat jako

$$a_{jj} = \frac{1}{(1 - r_j^2)\mathbf{x}_j'\mathbf{x}_j},$$

kde

$$r_j = r_{x_j.x_1x_2\cdots x_{j-1}x_{j+1}\cdots x_p}$$

je koeficient mnohonásobné korelace. Vysoký stupeň multikolinearity se projevuje vysokými hodnotami korelačních koeficientů r_j (blízkých 1), ale i vysokými hodnotami některých (nebo všech) jednoduchých korelačních koeficientů.

O multikolinearitě svědčí i **vysoké hodnoty poměru největšího a nejmenšího vlastního čísla**.

Důsledkem vysokých rozptylů odhadů jsou příliš dlouhé intervaly spolehlivosti, a tedy malá přesnost odhadu.

Logickým důsledkem multikolinearity je obtížné vyjádření individuálního vlivu jednotlivých vysvětlujících proměnných. Projeví se to **nízkými hodnotami** testových kritérií v testech t, nedovolujícími potvrdit závažnost jednotlivých regresorů v regresní funkci.

Někteří autoři doporučují **testovat hodnotu determinantu korelační matice R** vysvětlujících proměnných pomocí veličiny

$$W = -\left[n - 1 - \frac{1}{6}(2p + 7)\right] \ln |\mathbf{R}|,$$

která má při ortogonalitě proměnných rozdělení χ^2 s $p(p-1)/2$ stupni volnosti. **Jde o test hypotézy, že korelační matice je jednotková**.

Pro identifikaci proměnných způsobujících multikolinearitu se doporučují veličiny

$$F_j = \frac{n-p}{p-1}(d_{jj} - 1),$$

kde d_{jj} jsou diagonální prvky matice

$$\mathbf{D} = \mathbf{R}^{-1}.$$

Veličiny F_j mají v případech, kdy proměnná x_j nezpůsobuje multikolinearitu, rozdělení F s $p-1$ a $n-p$ stupni volnosti.

Poznamenejme, že přes značně nepříznivé důsledky se nemusí multikolinearita nepříznivě projevit na predikčních schopnostech regresního modelu.

Závažným důsledkem multikolinearity je značná výpočetní nespolehlivost a nestabilní hodnoty regresních odhadů. Stačí malý zásah do statistických údajů a výsledné odhady jsou odlišné.

C. Zlepšování podmíněnosti matice $\mathbf{X}'\mathbf{X}$ transformací proměnných:

Poznamenejme, že $\mathbf{X}$ je zde celá matice plánu.

Model centrovaných proměnných. Místo hodnot y_i a x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, používáme odchylky od aritmetických průměrů $\boxed{y_i - \bar{y}}$ a $\boxed{x_{ij} - \bar{x}_j.}$ Není obtížné ukázat, že hlavní předností cetrování proměnných je výpočetní zjednodušení. S výjimkou absolutního členu se odhady parametrů centrováním nezmění.

Model standardizovaných (normovaných) proměnných. Místo původních proměnných y_i a x_{ij} pracujeme s proměnnými ve tvaru

$$q_i = \frac{y_i - \bar{y}}{s_y} \quad \text{a} \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}},$$

kde s_y a s_{x_j} jsou směrodatné odchylky jednotlivých proměnných. Standardizací vysvětlujících proměnných dostáváme při použití metody nejmenších čtverců místo matice $\mathbf{X}'\mathbf{X}$ korelační matici

$$\mathbf{R} = \mathbf{Z}'\mathbf{Z}/n.$$

Vektor

$$\mathbf{Z}'\mathbf{q}/n$$

obsahuje jednoduché korelační koeficienty $\boxed{r_{yx_j}}$.

Standardizací proměnných se zmenšují zaokrouhlovací chyby a zlepšují se možnosti hodnocení individuálního vlivu proměnných pomocí regresních paramerů (viz tzv. beta koeficienty).

Model ortogonalizovaných vysvětlujících proměnných.

Pomocí **Grammova-Schmidtova ortogonalizačního postupu** a standardizací sloupcových vektorů získáme soustavu ortonormálních vektorů takových, že matice $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$. Užitečným důsledkem ortogonalizace je nezávislost regresních odhadů, tj.

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{I}_k$$

a tím i dobrá interpretace výsledků.

Model v kanonickém tvaru. Místo modelu ve tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pracujeme s modelem

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

kde matice

$$\mathbf{U} = \mathbf{X}\mathbf{V},$$

vektor

$$\boldsymbol{\gamma} = \mathbf{V}'\boldsymbol{\beta},$$

a $\mathbf{V}$ je matice standardizovaných vlastních vektorů odpovídajících vlastním číslům matice $\mathbf{X}'\mathbf{X}$. Odhady parametrů modelu v kanonickém tvaru:

$$\hat{\boldsymbol{\gamma}} = \mathbf{c} = \mathbf{L}^{-1}\mathbf{U}'\mathbf{Y},$$

kde $\mathbf{L}$ je diagonální matice s vlastními čísly matice $\mathbf{X}'\mathbf{X}$. Kovarianční matice odhadů

$$var(\mathbf{c}) = \sigma^2\mathbf{L}^{-1}$$

ukazuje, že i v tomto případě jsou odhady nezávislé. Rozptyly odhadů závisí na velikosti vlastních čísel (čím větší je vlastní číslo, tím menší je rozptyl odhadu). Vztah mezi původními a transformovanými odhady je možné vyjádřit jako

$$\mathbf{c} = \mathbf{V}'\mathbf{b}.$$

Reziduální součet čtverců se transformací nemění.

D. Hřebenová regrese.

Autoři hřebenové regrese vyšli ze skutečnosti, že při vysoké multikolinearitě jsou diagonální prvky matice $(\mathbf{X}'\mathbf{X})^{-1}$ příliš velké ($\mathbf{X}$ je zde celá matice plánu). Navrhli proto zkreslený odhad

$$\hat{\boldsymbol{\beta}}_H = (\mathbf{X}'\mathbf{X} + m\mathbf{I}_k)^{-1}\mathbf{X}'\mathbf{Y},$$

kde m je kladná konstanta. Pro $m = 0$ je $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}$. Zkreslení odhadu je

$$-m(\mathbf{X}'\mathbf{X} + m\mathbf{I}_k)^{-1}\boldsymbol{\beta}.$$

S růstem konstanty m se $\hat{\boldsymbol{\beta}}_H$ blíží k nule.

Při užití metody hřebenové regrese se většinou vychází z modelu v kanonickém tvaru

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Odhadem $\hat{\boldsymbol{\gamma}}$ hřebenou regresí je

$$\hat{\boldsymbol{\gamma}}_H = \mathbf{c}_H = (\mathbf{L} + m\mathbf{I}_k)^{-1}\mathbf{U}'\mathbf{Y},$$

kde $\mathbf{L}$ je diagonální matice s vlastními čísly matice $\mathbf{X}'\mathbf{X}$ a $\mathbf{U} = \mathbf{X}'\mathbf{V}$.

Vztah mezi odhady metodou nejmenších čtverců a hřebenovými odhady je možno zapsat jako

$$\hat{\boldsymbol{\beta}}_H = \mathbf{V}\mathbf{D}\mathbf{V}'\hat{\boldsymbol{\beta}},$$

kde $\mathbf{V}$ je matice vlastních vektorů odpovídajících $\mathbf{X}'\mathbf{X}$,
$\mathbf{D}$ je diagonální matice s prvky $\frac{\lambda_j}{\lambda_j+m}$ a
λ_j jsou vlastní čísla matice $\mathbf{X}'\mathbf{X}$.

Mezi odhady v původním a kanonickém tvaru je vztah

$$\mathbf{c}_H = \mathbf{V}'\hat{\boldsymbol{\beta}}_H.$$

Nechceme-li zvyšovat všechny diagonální prvky matice $\mathbf{X}'\mathbf{X}$ o stejnou konstantu (m) máme možnost použít metodu **zobecněné hřebenové regrese** , lišící se pouze v tom, že diagonální prvky se zvyšují o kladné konstanty m_j a $\boldsymbol{\gamma}$ se odhadujeme jako

$$c_{ZH} = (\mathbf{L} + \mathbf{M})^{-1}\mathbf{U}'\mathbf{Y},$$

kde $\mathbf{M}$ je diagonální matice s m_j na diagonále. Vztah mezi $\hat{\boldsymbol{\beta}}$ a $\hat{\boldsymbol{\beta}}_{ZH}$ lze opět vyjádřit

$$\hat{\boldsymbol{\beta}}_{\mathbf{ZH}} = \mathbf{VDV}'\hat{\boldsymbol{\beta}},$$

kde diagonální matice $\mathbf{D}$ obsahuje prvky $\frac{\lambda_j}{\lambda_j+m_j}$.

Problémem zůstává **určení konstanty** m, popř. konstant m_j. Platí tato skutečnost: pro určité hodnoty m se systém stabilizuje a dostává charakter ortogonálního systému a každá úloha má svou optimální hodnotu m.

Při použití metody hřebenové regrese bývá zvykem sledovat tzv. **HŘEBENOVOU STOPU**, tj. vztah mezi hodnotou určitého parametru a velikostí konstanty m.

V poslední době se doporučuje (pro grafické znázornění) používat místo m_j konstantu m_1 definovanou jako

$$m_1 = p - \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + m_j}.$$

Výhodou m_1 je okolnost, že pro $m_j \to \infty$ se blíží k p. Tato skutečnost umožňuje udělat si lepší představu o stabilizaci regresních odhadů při zvyšování hodnot m_j.

Některé výpočetní postupy používané v hřebenové regresi.

Při znalosti konstanty m (konstant m_j) můžeme na základě odhadů c_j, $j = 1, \ldots, p$, pro model v kanonickém tvaru snadno určit hřebenové odhady jako

$$c_{j,H} = \frac{c_j}{1 + m/\lambda_j}.$$

Ve skutečnosti však konstatu m neznáme, proto v původní práci autoři metody doporučili po převedení modelu do kanonického tvaru odhadnout m jako

$$k\hat{\sigma}^2 / \sum_{j=1}^{k} c_j^2.$$

Tento odhad m je možno dosadit do předchozího vzorce, a tak dostat "první" hřebenový odhad c_H. V postupu je třeba pokračovat s tím, že se tak podaří celý systém stabilizovat. Později došlo k rozpracování metody hřebenové regrese a objevila se celá řada možností odhadu konstanty m. Uveďme aspoň některé z nich:

(a) $\boxed{ps_\varepsilon^2 / \sum_{j=1}^{p} \hat{\beta}_j^2,}$ kde $\hat{\beta}_j$ jsou odhady metodou nejmenších čtverců modelu v původním tvaru a $s_\varepsilon^2 = \frac{SSE}{n-p-1}$

(b) $\boxed{ps_\varepsilon^2 / \sum_{j=1}^{k} \lambda_j c_j^2,}$ kde c_j jsou odhady parametrů modelu v kanonickém tvaru $c = \mathbf{V}'\hat{\beta}$ a $\mathbf{V}$ je matice charakteristických vektorů matice $\mathbf{X}'\mathbf{X}$.

(c) Hledání optimální hodnoty m z rovnice $\boxed{s_\varepsilon^2 \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j+m} = m \sum_{j=1}^{p} \frac{c_j^2\lambda_j^2}{\lambda_j+m^2}}$

(d) Hledání optimální hodnoty m z rovnice $\boxed{p = \frac{m}{s_\varepsilon^2} \sum_{j=1}^{p} \frac{c_j^2\lambda_j}{\lambda_j+m}}$

LITERATURA: Hebák, P., Hustopecký, J.(1987) Vícerozměrné statistické metody s aplikacemi, Praha, SNTL, Alfa

PŘÍKLAD 1: **Chemické složení portlandského cementu**

Portlandský cement je nejvíce používaným druhem cementu při výrobě betonu a malty. Obsahuje směs oxidů kovů alkalických zemin vápníku dále pak oxidy křemíku a hliníku.

Portlandský cement a podobné materiály jsou vyráběny pálením vápence (jako zdroje vápníku) s jílem nebo s pískem (zdroj křemíku), čímž vzniká slinek, ke kterému se v procesu mletí přidá sádrovec, jako regulátor tuhnutí. Výsledný prášek po smísení s vodou začne hydratovat a tím tuhne.

Portlandský cement byl poprvé vyroben ve Velké Británii na počátku 19. století a jeho název je odvozen od podobnosti s portlandským kamenem (stavební kámen), který se těží v Dorsetu na ostrově Isle of Portland, který leží v kanálu La Manche. Patent na tento cement získal britský zedník Joseph Aspdin v roce 1824.

Máme k dispozici údaje, které se týkají chemického složení portlandského cementu:

y	množství tepla v kaloriích na gram cementu
x_1	složka číslo 1 v % (Tricalcium aluminate 3CaO.Al2O3)
x_2	složka číslo 2 v % (Tricalciam silicate 3CaO.SiO2)
x_3	složka číslo 3 v % (Tetracalcium alumino ferrite 4CaO.Al2O3.Fe2O3)
x_4	složka číslo 4 v % (Dicalcium silicate 2CaO.SiO2)

Jde o velmi známá tzv. **Haldova data**, která byla publikována již v roce 1932 v Industrial and Engineering Chemistry 24.

Literatura: Woods, H., Steinour, H. H. and Starke, H. R. (1932) Effect of composition of Portland cement on heat evolved during hardening. *Industrial Engineering and Chemistry*, 24, 1207–1214.

Tato data byla převzata několika dalšími autory monografií o regresi, protože velmi dobře dokumentují některé problémy související s uplatněním regresní metody.

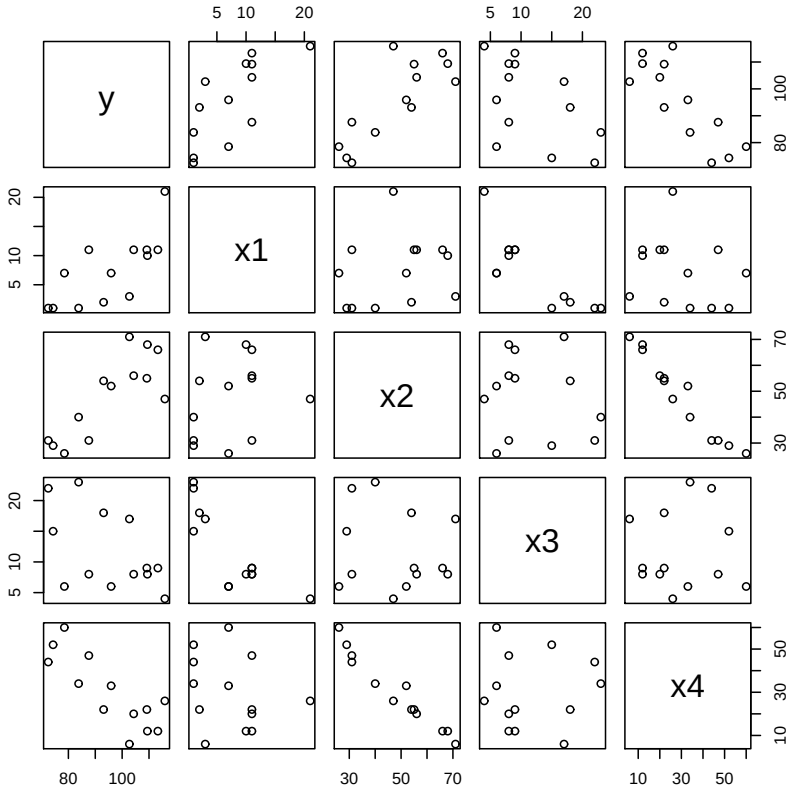
```
> cement <- matrix(c(78.5, 7, 26, 6, 60, 74.3, 1, 29, 15, 52, 104.3, 11,
+     56, 8, 20, 87.6, 11, 31, 8, 47, 95.9, 7, 52, 6, 33, 109.2, 11, 55,
+     9, 22, 102.7, 3, 71, 17, 6, 72.5, 1, 31, 22, 44, 93.1, 2, 54, 18,
+     22, 115.9, 21, 47, 4, 26, 83.8, 1, 40, 23, 34, 113.3, 11, 66, 9,
+     12, 109.4, 10, 68, 8, 12), ncol = 5, byrow = TRUE)
> colnames(cement) <- c("y", "x1", "x2", "x3", "x4")
```

```
> data <- data.frame(cement)
> print(data)
```

```
       y x1 x2 x3 x4
1   78.5  7 26  6 60
2   74.3  1 29 15 52
3  104.3 11 56  8 20
4   87.6 11 31  8 47
5   95.9  7 52  6 33
6  109.2 11 55  9 22
7  102.7  3 71 17  6
8   72.5  1 31 22 44
9   93.1  2 54 18 22
10 115.9 21 47  4 26
11  83.8  1 40 23 34
12 113.3 11 66  9 12
13 109.4 10 68  8 12
```

Data vykreslíme

```
> plot(data)
```



Obrázek 1: *Chemické složení portlandského cementu*

Předpokládejme model ve formě

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

```
> model <- lm(y ~ x1 + x2 + x3 + x4, data)
> summary(model)


Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared: 0.9824,        Adjusted R-squared: 0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Všimněme si, že podle statistiky F zamítáme nulovou hypotézu

$$H_0 : (\beta_1, \beta_2, \beta_3, \beta_4)' = (0, 0, 0, 0)' \quad vs \quad H_1 : (\beta_1, \beta_2, \beta_3, \beta_4)' \neq (0, 0, 0, 0)'$$

a model vysvětluje $97,36\%$ rozptylu (`Adjusted R-squared`). Přesto jednotlivé t–testy neoznačí ani jednu proměnnou jako statitisticky významnou. Tento paradox bývá důsledkem korelovaných regresorů x_1, x_2, x_3, x_4. Proto spočítáme nejprve korelační matici, následně její inverzi a všimneme si diagonálních prvků. Ještě si všimneme, že stejné výsledky dostaneme pomocí funkce `vif()` z knihovny `car`. Pro názornost hodnoty VIF vykreslíme.

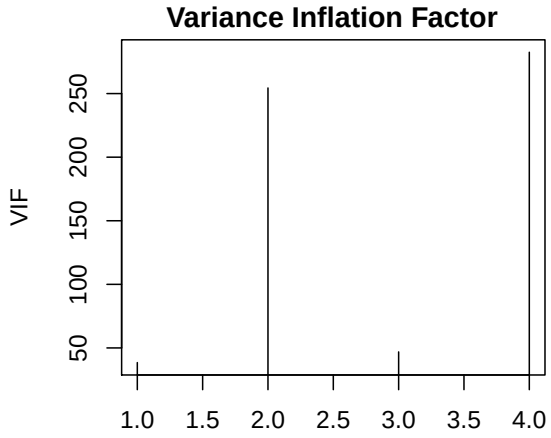
```
> print(Xcor <- cor(cement[, -1]))


           x1         x2          x3          x4
x1  1.0000000  0.2285795 -0.82413376 -0.24544511
x2  0.2285795  1.0000000 -0.13924238 -0.97295500
x3 -0.8241338 -0.1392424  1.00000000  0.02953700
x4 -0.2454451 -0.9729550  0.02953700  1.00000000


> print(Xcori <- solve(Xcor))


         x1       x2        x3       x4
x1 38.49621  94.1197  41.88410  99.7858
x2 94.11969 254.4232 105.09139 267.5394
x3 41.88410 105.0914  46.86839 111.1451
x4 99.78580 267.5394 111.14509 282.5129


> print(VIF <- diag(Xcori))
```

```
      x1        x2        x3        x4
38.49621 254.42317  46.86839 282.51286
```

```
> library(car)
> vif(model)
```

```
      x1        x2        x3        x4
38.49621 254.42317  46.86839 282.51286
```

```
> par(mar = c(2, 5, 1.5, 0.5) + 0.05)
> plot((1:length(VIF)), VIF, type = "h", main = "Variance Inflation Factor")
```



Obrázek 2: VIF pro *Chemické složení portlandského cementu*

Všechny VIF značně přesahují hodnotu 10, uvedenou už jako příliš velkou.

Všimněme si dále vlastních čísel korelační matice.

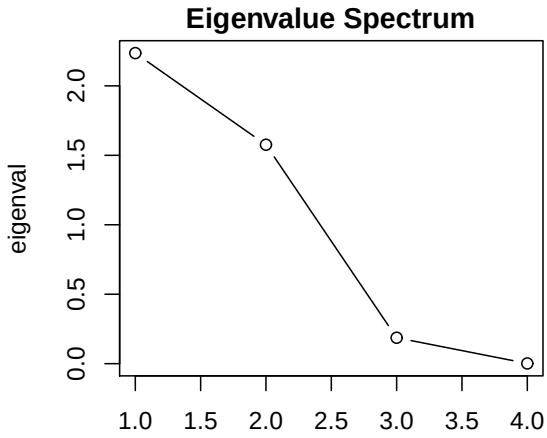
```
> X.eig <- eigen(Xcor)
> eigenval <- X.eig$values
> eigenvec <- X.eig$vectors
> print(conditionnumber <- max(eigenval)/min(eigenval))
```

```
[1] 1376.881
```

```
> print(conditionindex <- max(eigenval)/eigenval)
```

```
[1]    1.000000    1.418534   11.980870 1376.880621
```

```
> par(mar = c(2, 5, 1.5, 0.5) + 0.05)
> plot((1:length(eigenval)), eigenval, type = "b", main = "Eigenvalue Spectrum")
```



Obrázek 3: Spektrální rozklad korelační matice dat *Chemické složení portlandského cementu*

Testujme hodnotu determinantu korelační matice $\mathbf{R}$ vysvětlujících proměnných pomocí veličiny

$$W = -\left[n - 1 - \frac{1}{6}(2p + 7)\right] \ln |\mathbf{R}|,$$

která má při ortogonalitě proměnných rozdělení χ^2 s $p(p-1)/2$ stupni volnosti. **Jde o test hypotézy, že korelační matice je jednotková**.

```
> n <- nrow(data)
> p <- 4
> const <- -(n - 1 - (2 * p + 7)/6)
> alpha <- 0.05
> print(W <- const * log(det(Xcor)))
```

```
[1] 65.00172
```

```
> dg <- p * (p - 1)/2
> print(chi2Kvantily <- qchisq(alpha, dg))
```

```
[1] 1.635383
```

Protože statistika W je větší než příslušný kvantil χ^2 rozdělení, považujeme multikolinearitu za prokázanou.

Pro identifikaci proměnných způsobujících multikolinearitu se doporučují veličiny

$$F_j = \frac{n-p}{p-1}(d_{jj} - 1),$$

kde d_{jj} jsou VIF prvky, tj. diagonální prvky matice $\mathbf{D} = \mathbf{R}^{-1}$. Veličiny F_j mají v případech, kdy proměnná x_j nezpůsobuje multikolinearitu, rozdělení F s $k-1$ a $n-p$ stupni volnosti.

```
> const <- (n - p)/(p - 1)
> print(Fj <- const * (VIF - 1))
```

```
      x1       x2       x3       x4
112.4886 760.2695 137.6052 844.5386
```

```
> print(Fkvantily <- qf(alpha, p - 1, n - p))
```

```
[1] 0.1134778
```

Vidíme, že všechny hodnoty F_j jsou vysoce významné. Poněkud větší je vliv dvojice x_2 a x_4.

E. Autokorelace reziduí

V regresních modelech pro časové řady je třeba věnovat velkou pozornost problematice autokorelovaných reziduí. Ve většině případů se u časových řad s autokorelací reziduí setkáme, nebot' hodnota pozorování v časovém okažiku t velmi pravděpodobně ovlivní následující hodnoty.

Pro testování autokorelace reziduí prvního řadu je používán Durbin–Watsonův test

Durbin–Watsonův test autokorelace reziduí 1. řádu

Durbin-Watsonova statistika je definována vztahem

$$D = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}.$$

Protože platí $(a-b)^2 \leq 2a^2 + 2b^2$, dostáváme

$$D \leq \frac{2\sum_{i=2}^{n} r_i^2 + 2\sum_{i=2}^{n} r_{i-1}^2}{\sum_{i=1}^{n} r_i^2} \leq 4 \qquad \Rightarrow \qquad \boxed{0 \leq D \leq 4}.$$

Vzhledem k tomu, že $Er = 0$, bude pro větší hodnoty n platit

$$\sum_{i=2}^{n} r_i^2 \doteq \sum_{i=1}^{n} r_i^2 \doteq \sum_{i=1}^{n-1} r_{i+1}^2.$$

Označme výběrový autokorelační koeficient:

$$\hat{\rho}(1) = \frac{\widehat{E}(r_i r_{i+1})}{\sqrt{\widehat{D}r_i \widehat{D}r_{i+1}}} = \frac{\sum_{i=1}^{n-1} r_{i+1} r_i}{\sqrt{\sum_{i=1}^{n-1} r_i^2 \sum_{i=1}^{n-1} r_{i+1}^2}} \Rightarrow D \approx 2(1-\hat{\rho}_1) \quad \text{nebo} \quad \hat{\rho}(1) \approx 1 - \tfrac{D}{2}.$$

Pokud budou **rezidua málo korelovaná**, hodnota D se bude pohybovat **kolem 2**.

Kladná korelace způsobí, že $D \in (0, 2)$ a **záporná korelace** způsobí, že $D \in (2, 4)$.

Přesné rozdělení statistiky D závisí na tvaru matice plánu $\mathbf{X}$, proto jsou tabelovány intervaly d_L a d_U, ve kterých se nachází kritické hodnoty (pro různá n, k a α).

Dolní a horní hranice Durbin-Watsonova testu na 5% hladině významnosti										
	k=1		k=2		k=3		k=4		k=5+	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
100+	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

kde k je počet nezávisle proměnných v regresní rovnici.

Pro **rychlé posouzení autokorelace prvního řadu** vystačíme s následující tabulkou:

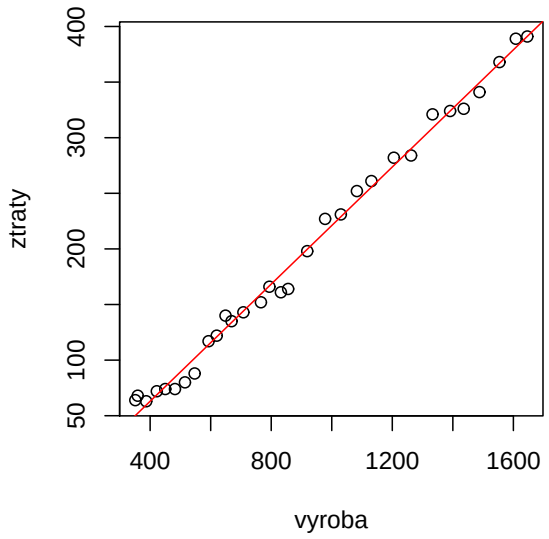
Pokud hodnota Durbin-Watsonovy statistiky D bude v mezích				
0 až d_L	d_L až d_U	d_U až $(4 - d_U)$	$(4 - d_U)$ až $(4 - d_L)$	$(4 - d_L)$ až 4
Zamítáme H_0 KLADNÁ autoko-relace	Ani nezamítáme ani nepřijímáme H_0	Nezamítáme nulovou hypotézu H_0	Ani nezamítáme ani nepřijímáme H_0	Zamítáme H_0 NEGATIVNÍ auto-korelace

V knihovně `lmtest` prostředí R je Durbin–Watsonův test uveden jako funkce `dwtest()`.

Příklad 1: **Ztráty vyrobené vody v letech 1953–1983**

Podle *Historické statistické ročenky* z roku 1985 se sledovaly ztráty vyrobné vody (zjištěné jako podíl mezi možstvím vyrobené a fakturované vody) jako funkci množství vyrobené vody.

```
> vyroba <- c(351, 359, 387, 422, 450, 482, 515, 547, 593, 620, 649, 669,
+     708, 766, 794, 832, 856, 919, 978, 1030, 1083, 1131, 1205, 1262,
+     1333, 1391, 1436, 1488, 1554, 1608, 1646)
> ztraty <- c(64, 68, 63, 72, 74, 74, 80, 88, 117, 122, 140, 135, 143,
+     152, 166, 161, 164, 198, 227, 231, 252, 261, 282, 284, 321, 324,
+     326, 341, 368, 389, 391)
> par(mar = c(5, 5, 1.5, 0.5) + 0.05)
> plot(ztraty ~ vyroba)
> abline(lm(ztraty ~ vyroba), col = "red")
```

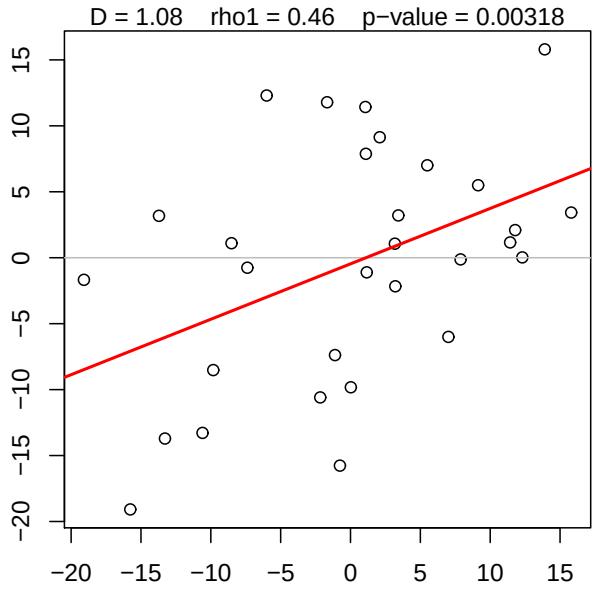


Obrázek 4: *Ztráty vyrobené vody v letech 1953–1983*

```
> model <- lm(ztraty ~ vyroba)
> summary(model)


Call:
lm(formula = ztraty ~ vyroba)

Residuals:
    Min      1Q  Median      3Q     Max
-19.084  -6.693   1.066   6.252  15.799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.340155   4.109317  -10.30 3.34e-11 ***
vyroba        0.263346   0.004151   63.45  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.261 on 29 degrees of freedom
Multiple R-squared: 0.9928,        Adjusted R-squared: 0.9926
F-statistic:  4025 on 1 and 29 DF,  p-value: < 2.2e-16


> library(lmtest)
> (DWtest <- dwtest(ztraty ~ vyroba, alternative = "two.sided"))


        Durbin-Watson test

data:  ztraty ~ vyroba
DW = 1.0819, p-value = 0.003179
alternative hypothesis: true autocorelation is not 0
```

Protože p–hodnota je menší než 0,05, zamítáme nulovou hypotézu, že data nejsou korelovaná.

```
> n <- length(ztraty)
> x <- resid(model)[1:(n - 1)]
> y <- resid(model)[2:n]
> par(mfrow = c(1, 1), mar = c(2, 2, 1, 0) + 0.05)
> txt <- paste("D =", round(DWtest$statistic, 2), "   rho1 =", round(1 -
+     0.5 * DWtest$statistic, 2), "   p-value =", round(DWtest$p.value,
+     5))
> plot(x, y)
> abline(h = 0, col = "gray")
> abline(lm(y ~ x), col = "red", lwd = 2)
> mtext(txt)
```



Obrázek 5: Grafické testování autokorelace *Ztráty vyrobené vody v letech 1953–1983*

F. Úkol: *Canadian Women's Labour-Force Participation Data*

Načtěte informační a datový soubor `women.inf` a `women.txt` . Jako nezávisle proměnnou uvažujte rok. Pro dvě časové řady [2] *Percent of adult women in the workforce* a [4] *Men's average weekly wages, in constant 1935 dollars and adjusted for current tax rates* nalezněte vhodnou trendovou funkci a následně proveďte analýzu reziduí. Zaměřte se především na autokorelaci.