6 Geometrie metody nejmenších čtverců a regresní diagnostika

Průvodce studiem

I tato kapitola je velmi důležitá pro pochopení principů statistické analýzy vícerozměrných dat. Počítejte nejméně se čtyřmi hodinami usilovného studia s tím, že se k probírané látce budete podle potřeby ještě vracet po pochopení dalších souvislostí.

6.1 Geometrie metody nejmenších čtverců

Uvažujeme klasický lineární regresní model

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Jak víme z předchozí kapitoly, odhad parametrů můžeme vyjádřit explicitně

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Vektor $\hat{\mathbf{y}} = \mathbf{Xb}$ je lineární kombinací vektorů regresorů, tj. leží v prostoru (přímce, rovině, nadrovině), jehož dimenze je rovna počtu regresorů. Dosadíme-li za $\mathbf{b}$, dostaneme

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{Hy}$$

Matice $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ je matice *projekce* vektoru $\mathbf{y}$ do prostoru určeného vektory regresorů. Požadavek formulovaný v metodě nejmenšich čtverců, tj. RSS $= (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$ vlastně znamená, že tato projekce je ortogonální. Pak tedy vektory $\hat{\mathbf{y}}$ a $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ jsou ortogonální vektory, tzn. $\hat{\mathbf{y}}^T\ \mathbf{e} = \mathbf{e}^T\ \hat{\mathbf{y}} = 0$, o čemž se velmi snadno můžeme přesvědčit:

$$(\mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb}) = \mathbf{b}^T\mathbf{X}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{b}^T(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{Xb}) = 0,$$

neboť výraz v poslední závorce je nulový vektor, viz normální rovnice (9).

Vektoru $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ se říká vektor residuí, jeho složkám $e_i = y_i - \hat{y}_i$ pak *residua*. Součet a tedy i průměr residuí je roven nule:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0,$$

neboť z první normální rovnice platí, že $\bar{y} = \mathbf{b}^T\bar{\mathbf{x}}$, kde $\bar{\mathbf{x}}^T = [1, \bar{x_1}, \bar{x_2}, \ldots, \bar{x_k}]$, tudíž

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}\hat{y}_i - \sum_{i=1}^{n}y_i = \mathbf{b}^T\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}) = 0,$$

neboť součet odchylek od průměru je nulový.

6.2 Rozklad součtu čtverců

Variabilitu vysvětlované veličiny můžeme vyjádřit jako součet čtverců odchylek pozorovaných hodnot od jejich průměru. Tuto charakteristiku nazýváme celkový součet čtverců, TSS.

$$\mathrm{TSS} = (\mathbf{y} - \bar{\mathbf{y}})^T(\mathbf{y} - \bar{\mathbf{y}})$$

Lze ukázat, že tuto celkovou sumu čtverců můžeme rozložit na dvě složky

$$\mathrm{MSS} = (\hat{\mathbf{y}} - \bar{\mathbf{y}})^T(\hat{\mathbf{y}} - \bar{\mathbf{y}})$$

a už dříve definovanou

$$\mathrm{RSS} = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{e}^T\mathbf{e}$$

Platí tedy, že

$$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS},$$

MSS je ta část z celkového součtu čtverců, která je vysvětlena závislostí vysvětlované veličiny na regresorech, zbylou část (RSS) lineární závislostí vysvětlit nelze.

Nyní můžeme zavést důležitou charakteristiku toho, jak úspěšně regresní model vysvětluje variabilitu vysvětlované veličiny. Této charakteristice se říká *koeficient (index) determinace*, R^2.

$$R^2 = \frac{\mathrm{MSS}}{\mathrm{TSS}} = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Vidíme, že $0 \leq R^2 \leq 1$. Hodnota indexu determinace $R^2 = 1$, když $\mathrm{RSS} = 0$, tzn. regresní model vysvětluje závislost vysvětlované veličiny na regresorech úplně (dokonalá lineární závislost). Naopak, $R^2 = 0$, když model nevysvětluje nic, tedy $\mathrm{RSS} = \mathrm{TSS}$, což nastane jen tehdy, když všechny odhady $b_1 = b_2 = \ldots = b_k = 0$ a $b_0 = \bar{y}$, např. pro $k = 1$ je regresní přímka rovnoběžná s osou x v úrovni $b_0 = \bar{y}$.

Z rozkladu celkového součtu čtverců vychází i analýza rozptylu, která je obvyklou součástí regresních programů. Tabulka analýzy rozptylu má většinou tento formát:

zdroj variability	stupně volnosti	součet čtverců	průměrný čtverec	F	$p-$value
model	k	MSS	MSS$/k$	$\frac{\text{MSS}/k}{\text{RSS}/(n-k-1)}$	$0.\ldots$
error	$n-k-1$	RSS	RSS$/(n-k-1)$		
total	$n-1$	TSS			

Za předpokladu, že průměrný čtverec $s^2 = \text{RSS}/(n-k-1)$ je opravdu nestranným odhadem rozptylu náhodné složky (σ^2), tzn. v modelu jsou zařazeny všechny relevantní regresory a RSS není zvětšeno systematickou závislostí na nezařazeném regresoru (podrobněji viz např. Draper a Smith, kap. 2 a 24) a náhodné kolísání má normální rozdělení, má statistika F rozdělení $F \sim F_{k,n-k-1}$ a můžeme ji užít k testu hypotézy

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ proti

H_1 : aspoň jeden parametr $\beta_j \neq 0, \qquad j = 1, 2, \ldots, k$

Povšimněme si, že důležitou informaci o variabilitě residuí $e_i = y_i - \bar{y}_i$ a tím i o shodě modelem predikovaných hodnot $\bar{y}_i$ s pozorovanými hodnotami y_i nám poskytuje směrodatná odchylka residuí (square root mean error)

$$s = \sqrt{\frac{\text{RSS}}{n-k-1}}$$

Index determinace má tendenci nadhodnocovat podíl modelu na vysvětlení celkové variability veličiny y, mimo jiné i proto, že kvůli náhodnému kolísání jsou odhady $b_j \neq 0$ i tehdy, když $\beta_j = 0, \quad j = 1, 2, \ldots, k$. Proto se zavádí tzv. adjustovaný index determinace R^2_{adj},

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

Vidíme, že $R^2_{adj} < R^2$, rozdíl je výrazný tehdy, když počet pozorování je jen o málo větší než počet regresorů v modelu. Naopak hodnota R^2_{adj} se přibližuje R^2 pro $n \gg k$.

6.3 Regresní diagnostika

Další informace o vhodnosti modelu a o tom, zda jsou splněny předpoklady učiněné pro klasický lineární model můžeme získat z analýzy residuí. Vektor residuí můžeme vyjádřit pomocí projekční matice $\mathbf{H}$:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{I}\mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Pak kovarianční matice residuí je

$$\begin{gathered}\mathrm{cov}(\mathbf{e}) = \mathrm{cov}\left[(\mathbf{I}-\mathbf{H})\mathbf{y}\right] = (\mathbf{I}-\mathbf{H})\mathrm{cov}(\mathbf{y})(\mathbf{I}-\mathbf{H})^T = \\ (\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{H})^T = \sigma^2\mathbf{I}(\mathbf{I}-\mathbf{H})^T = \\ \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^T = \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}^T+\mathbf{H}\mathbf{H}^T) = \\ \sigma^2(\mathbf{I}-\mathbf{H})\end{gathered}$$

neboť projekční matice $\mathbf{H}$ je symetrická ($\mathbf{H}^T = \mathbf{H}$) a idempotentní ($\mathbf{H}^2 = \mathbf{H}$):

$$\mathbf{H}\mathbf{H}^T = \mathbf{H}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

Poznámka – vektor reziduí $\mathbf{e}$ je náhodný vektor (je to výraz, ve kterém jsou náhodné vektory $\mathbf{y}$ a $\mathbf{b}$).

Matice $\mathbf{H}$ s prvky $h_{ij}, \quad i,j = 1,2,\ldots,n$ je symetrická, ale nemusí být diagonální. Jak bylo v předchozím odstavci ukázáno, kovarianční matice vektoru residuí je rovna

$$\mathrm{cov}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$$

Nestranným odhadem parametru σ^2 je reziduální rozptyl (tzn. rozptyl ε_i):

$$s^2 = \frac{1}{n-k-1}\,\mathbf{e}^T\mathbf{e}$$

Dále uvedeme některé charakteristiky, které se užívají v tzv. *regresní diagnostice*, tj. při analýze vhodnosti modelu.

Klasická residua
Jsou to residua, který už jsme se zabývali,

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}.$$

Jejich rozptyly

$$\mathrm{var}(e_i) = s_e^2(1-h_{ii}),$$

nejsou konstantní, i když $\mathrm{var}(\epsilon_i) = \sigma^2$ konstantní je.

Normovaná residua
Jsou to klasická residua, vydělená reziduální směrodatnou odchylkou:

$$e_{Ni} = \frac{e_i}{s}$$

Jejich rozptyl je roven

$$\mathrm{var}(e_{Ni}) = 1 - h_{ii},$$

tedy nemusí být roven jedné.

Standardizovaná rezidua

Někdy se jim říká vnitřně studentizovaná residua (internally studentized), jsou definována takto:

$$e_{Si} = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

a jejich rozptyl je kostantní, roven jedné.

Plně studentizovaná rezidua

Podle techniky užité v jejich definici se jim říká také JACKKNIFE residua, jsou konstruována tak, že vždy pro $i-$tý bod se residuum počítá z modelu, jehož parametry byly odhadnuty ze zbývajících $n-1$ bodů, tedy vždy $i-$tý bod se vypustí.

$$e_{Ji} = \frac{e_{(-i)}}{s_{(-i)}\sqrt{1-h_{ii}}}.$$

kde $s_{(-i)}$ je residuální směrodatná odchylka při vynechání i-tého bodu (řádku datové matice). Tato residua mají $t-$rozdělení, $e_{Ji} \sim t(n-k-2)$.

Leverage

Tyto charakteristiky ohodnocují vliv i-tého bodu na hodnoty odhadů parametrů. Jsou to diagonální prvky projekční matice, tedy hodnoty h_{ii}. Platí, že

$$0 < h_{ii} < 1 \qquad \text{a} \qquad \sum_{i=1}^{n} h_{ii} = k+1,$$

kde k je počet regresorů. Hodnota h_{ii} je úměrná vzdálenosti i-tého pozorování od těžiště (v k-rozměrném prostoru regresorů), h_{ii} se považuje za velké, když h_{ii} je větší než dvojnásobek průměrné hodnoty, tj. $h_{ii} > 2(k+1)/n$).

Cookova vzdálenost

Tato charakteristika slouží také k posouzení vlivu i-tého pozorování na odhady parametrů modelu, tj. na hodnoty $\mathbf{b}$. Je to vlastně relativní změna reziduálního součtu čtverců způsobená vypuštěním i-tého pozorování. Cookova vzdálenost pro i-té pozorování je definována

$$C_i = \frac{(\mathbf{y}-\hat{\mathbf{y}}_{(-i)})^T(\mathbf{y}-\hat{\mathbf{y}}_{(-i)})}{ps^2} = \frac{(\mathbf{b}-\mathbf{b}_{(-i)})^T(X^TX)(\mathbf{b}-\mathbf{b}_{(-\mathbf{i})})}{ps^2} = \frac{h_{ii}}{p(1-h_{ii})}e_{Si}^2$$

kde $\mathbf{b}_{(-i)}$ jsou jsou jackknife odhady (spočítané při vypuštění i-tého bodu) a p je počet odhadovaných parametrů. Cookova vzdálenost ohodnocuje vliv i-tého pozorování na odhad vektoru regresních parametrů $\mathbf{b}$. Je-li Cookova vzdálenost $C_i \geq 1$, i-pozorování velmi podstatně ovlivňuje odhady parametrů.

6.4 Autokorelace

Při posuzování předpokladu o nekorelovanosti residuí se obvykle vychází modelu autokorelačního procesu prvního řádu – AR(1):

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + u_i \qquad \text{kde} \qquad u_i \sim N(0, \sigma^2)$$

Autokorelační koeficient prvního řádu ρ_1 odhadujeme jako

$$\hat{\rho}_1 = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$$

K testování korelovanosti residuí se pak užívá *Waldův test*

$$W_a = \frac{n\hat{\rho}_1^2}{1 - \hat{\rho}_1^2} \sim \chi^2(1)$$

nebo ve statistickém software běžně implementovaná *Durbin – Watsonova* statistika

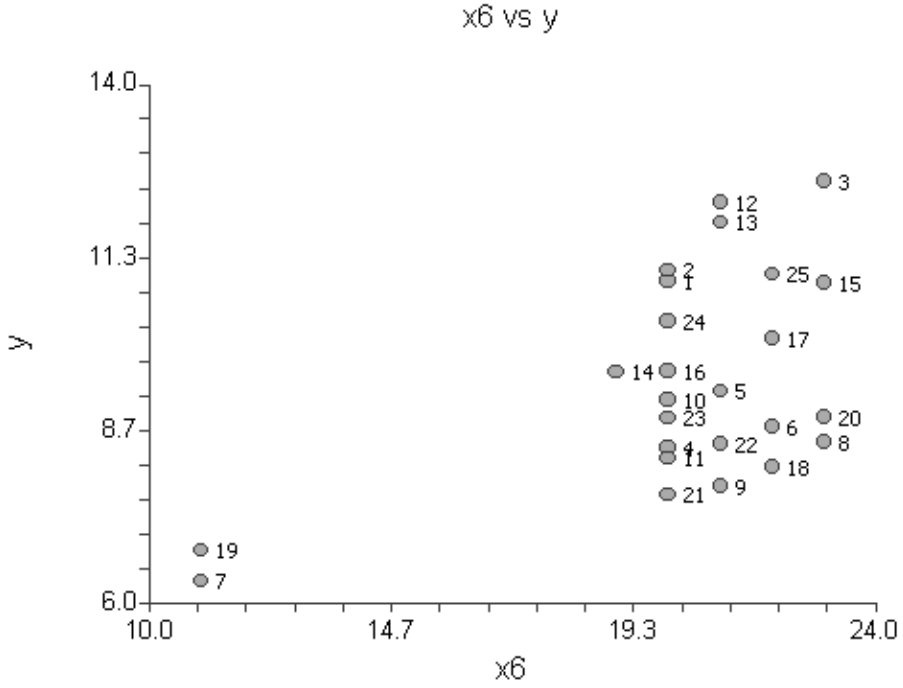
$$D_W = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \simeq 2\,(1 - \hat{\rho}_1)$$

Pro tuto statistiku platí $0 \leq D_W \leq 4$, $E(D_W) = 2$ při $\rho_1 = 0$. Kvantily této statistiky je obtížné vyjádřit explicitně, proto pro Durbin–Watsonův test statistické programy neposkytují u jiných testů obvyklý komfort, totiž i dosaženou významnost (p). Při rozhodování je pro hodnoty statistiky velmi blízké dvěma spoléhat na intuici a považovat residua za nekorelované. Pro serióznější úsudek lze využít přibližné kritické hodnoty, které jsou tabelovány, např. [16].

Příklad 6.1 Data pro tento příklad jsou převzata z knihy [8], příklad 01A, a jsou uvedena i v následující tabulce. Veličina y je měsíční spotřeba páry ve firmě, veličina $x6$ je počet pracovních dní v měsíci a veličina $x8$ je venkovní teplota ve stupních Fahrenheita. Úloha, kterou máme řešit, je odhadnout parametry lineárního regresního modelu a posoudit, zda je tento model vhodný pro vysvětlení závislosti y na $x6$ a $x8$. V řešení tohoto příkladu byl užit statistický programový systém NCSS [14], zejména modul Multiple Regression, old version. Výstupy uvádíme bez větších editačních úprav v surovém stavu.

i	y	$x6$	$x8$	i	y	$x6$	$x8$
1	10.98	20	35.3	14	9.57	19	39.1
2	11.13	20	29.7	15	10.94	23	46.8
3	12.51	23	30.8	16	9.58	20	48.5
4	8.40	20	58.8	17	10.09	22	59.3
5	9.27	21	61.4	18	8.11	22	70.0
6	8.73	22	71.3	19	6.83	11	70.0
7	6.36	11	74.4	20	8.88	23	74.5
8	8.50	23	76.7	21	7.68	20	72.1
9	7.82	21	70.7	22	8.47	21	58.1
10	9.14	20	57.5	23	8.86	20	44.6
11	8.24	20	46.4	24	10.36	20	33.4
12	12.19	21	28.9	25	11.08	22	28.6
13	11.88	21	28.1				

Vyšetřujeme-li nějakou závislost, vždy je dobré data nejdříve prohlédnout jednoduchými prostředky popisné statistiky. Ty nám často pomohou odhalit zajímavé věci v datech, např. odlehlé hodnoty. To můžeme vidět i na následujícím grafu, kde pozorování na řádcích 7 a 19 jsou zcela mimo hodnoty ostatních pozorování (patrně je to počet pracovních dnů v měsících, kdy byly dovolené a ve firmě se nepracovalo). Tato pozorování jsou apriori podezřelá a při diagnostice modelu je nutné si na ně dát pozor.



Další užitečné nahlédnutí do analyzovaných dat je výběrová korelační matice, která je volitelnou součástí výstupu z modulu Multiple Regression:

```
Multiple Regression Report
Dependent   y
Correlation Matrix Section
             x6         x8         y
x6          1.000000 -0.209761  0.536122
x8         -0.209761  1.000000 -0.845244
y           0.536122 -0.845244  1.000000
```

Vidíme, že regresory $x6$ a $x8$ jsou jen slabě korelovány (korelační koeficient je -0,21), takže matice regresorů má plnou hodnost, nehrozí numerické potíže spojené se špatnou podmíněností.

Další pro nás důležitou součástí výstupu z modulu Multiple Regression je Regression Equation Section, kde jsou odhady parametrů modelu.

```
Regression Equation Section
Independent   Regression     Standard     T-Value      Prob
Variable      Coefficient    Error      (Ho: B=0)     Level
Intercept      9.12688       1.102801       8.2761    0.000000
x6             0.2028154     4.576761E-02   4.4314    0.000210
x8            -7.239294E-02  7.999381E-03   9.0498    0.000000
```

Vidíme, že u všech tří odhadovaných parametrů zamítáme nulovou hypotézu. Model tedy neobsahuje nadbytečné parametry.

Jak vidíme v následující tabulce ANOVA, model vysvětluje významnou část z celkové variability veličiny y, podle hodnoty $R^2 = 0.8491$ asi 85% z celkové variability. Odhad residuální směrodatné odchylky je uveden jako Root Mean Square Error a je roven přibližně 0,66. Druhá mocnina této charakteristiky je pak odhadem residuálního rozptylu σ^2.

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	2220.294	2220.294		
Model	2	54.1871	27.09355	61.9043	0.000000
Error	22	9.628704	0.4376684		

Root Mean Square Error	0.6615651
Mean of Dependent	9.424
R-Squared	0.8491
Adj R-Squared	0.8354

Pro posouzení vhodnosti modelu jsou důležité výstupy z regresní diagnostiky. Durbin-Watsonova statistika je velmi blízká hodnotě 2, tak se nemusíme znepokojovat autokorelací residuí.

Durbin-Watson Value	2.1955

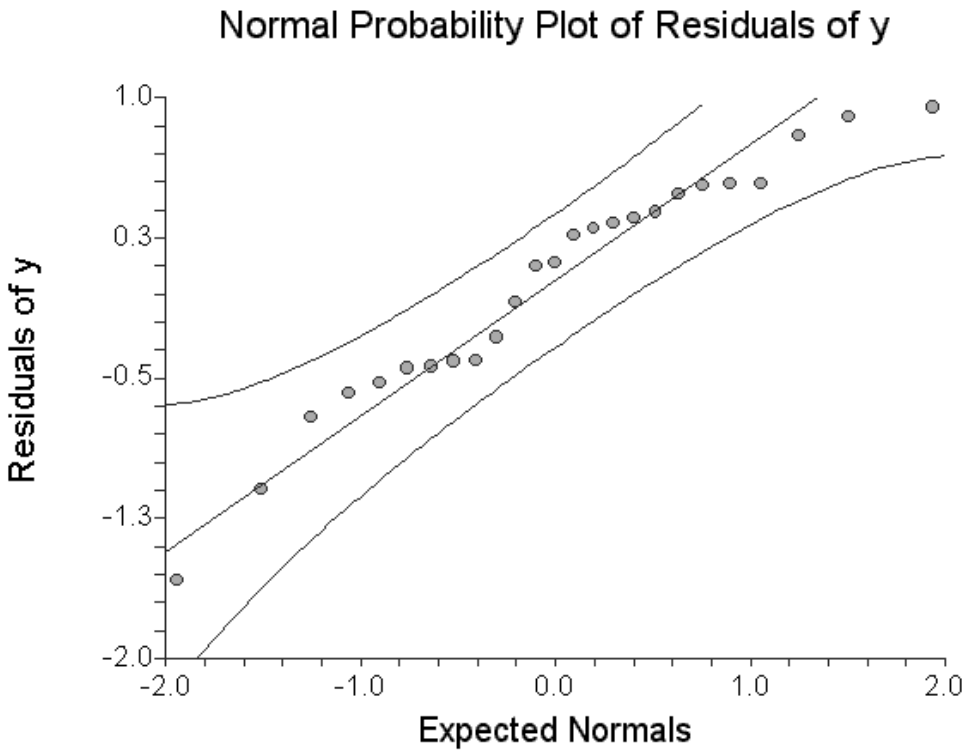
V následující tabulce jsou jackknife residua označena jako Rstudent. Pozornost věnujme především řádkům 7 a 19, které byly podezřelé už při předběžné jednoduché analýze a pak těm řádkům, kde studentizovaná residua jsou v absolutní hodnotě velká (zhruba větší než 2, což je přibližná hodnota kvantilu $t_{n-k-1}(0.975)$).

Regression Diagnostics Section

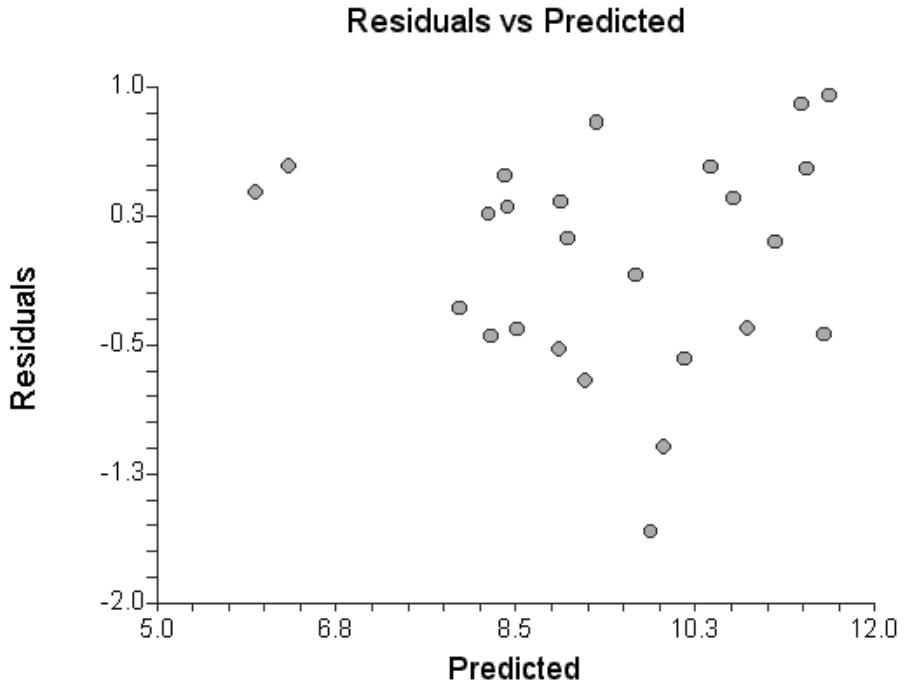
Row	Studentized Residual	Rstudent	Hat Diagonal	Cook's D
1	0.556825	0.547897	0.085491	0.009662
2	0.156002	0.152500	0.118877	0.001094
3	1.531855	1.583464	0.124826	0.111564
4	-0.814515	-0.808066	0.045374	0.010511
5	0.511834	0.503070	0.056434	0.005223
6	0.487210	0.478597	0.117502	0.010535
7	0.789332	0.782341	0.447410	0.168151
8	0.436764	0.428584	0.184719	0.014407
9	-0.711757	-0.703540	0.095491	0.017827
10	0.184529	0.180426	0.043373	0.000515
11	-2.452153	-2.810441	0.046418	0.097567
12	1.442820	1.481481	0.118566	0.093341
13	0.853098	0.847621	0.123991	0.034337
14	-0.913679	-0.910106	0.079880	0.024158
15	0.843302	0.837562	0.075758	0.019431
16	-0.142369	-0.139160	0.043079	0.000304
17	1.241672	1.258005	0.065527	0.036036
18	-0.658977	-0.650276	0.109838	0.017861
19	1.086621	1.091328	0.436460	0.304828
20	0.798049	0.791237	0.167792	0.042803
21	-0.450525	-0.442212	0.094228	0.007039
22	-1.100280	-1.105840	0.048654	0.020638
23	-1.697613	-1.779205	0.050307	0.050886
24	-0.644223	-0.635433	0.095790	0.014656
25	-0.708085	-0.699826	0.124217	0.023705

Největší residuum v absolutní hodnotě má pozorování 11, ale jak vidíme z ostatních statistik, neovlivňuje nijak významně hodnoty odhadů, je to odlehlý bod, který zvětšuje residuální rozptyl. Statistiky h_{ii} mají největší pozorování 7 a 19, jejich hodnoty jako jediné přesahují mezní hodnotu $2p/n = 6/25$, mají i největší Cookovu vzdálenost, ale zdaleka nedosahující hodnotu 1. Tyto dva body jsou tzv. vlivné, ale nevybočující. Naopak přispívají ke snížení residuálního rozptylu.

Následující QQ graf residuí ukazuje, že rozdělení residuí můžeme považovat za normální, odchylky od přímky nejsou velké. Tedy data i model vyhovují předpokladu (5) klasického modelu a výsledky testů o parametrech modelu můžeme považovat za spolehlivé (nezavádějící).



Graf residuí proti odhadovaným hodnotám $\hat{y}$ ukazuje, že residuální rozptyl můžeme považovat za konstatní, „kazí“ to jen bod 11 s residuem menším než -2.



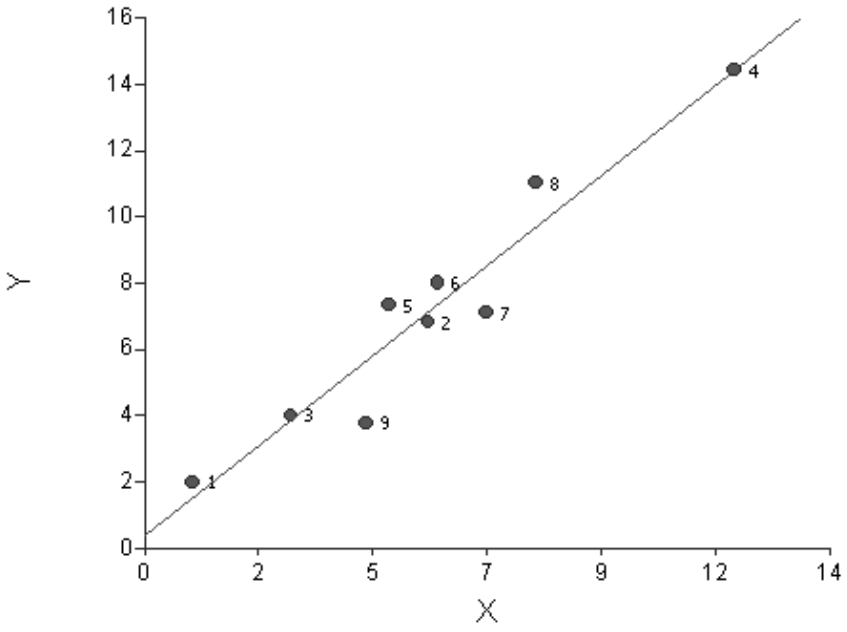
Výsledky tedy můžeme shrnout takto:

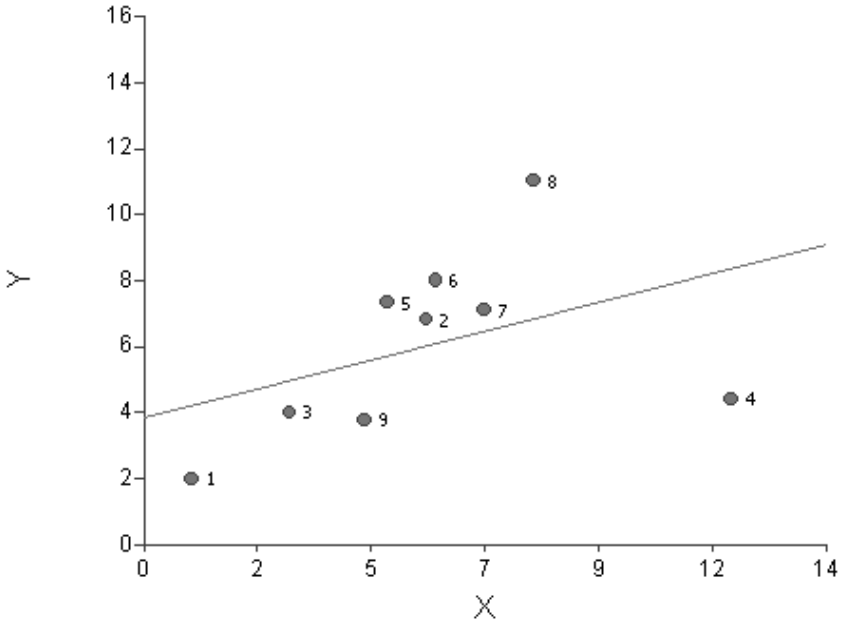
Hodnoty veličiny y (měsíční spotřeby páry ve firmě) lze uspokojivě odhadovat lineární závislostí na veličině $x6$ (počet pracovních dní v měsíci) a veličiny $x8$ (vnější teplota). Očekávanou spotřebu páry lze vyjádřit vztahem

$$\hat{y} = 9.127 + \ 0.2028\ x6 - 0.07239\ x8.$$

Směrodatná odchylka předpovědi je 0.662, model vysvětluje 85% z celkové variability spotřeby páry.

Příklad 6.2 Význam statistik pro diagnostiku ukazuje jednoduchý příklad modelu s jedním regresorem. Na obrázcích jsou data a proložené regresní přímky.





Odlišnost je jen v hodnotách vertikální souřadnice bodu 4. V obou úlohách mají body 4 velké hodnoty h_{ii}, ale liší se v hodnotách ostaních diagnostických statistik.

V případě prvním má bod 4 malé jackknife residuum i Cookovu vzdálenost, tzn. jeho vypuštěním či přidáním se hodnoty odhadů příliš nemění:

```
Row  Residual  Rstudent  Diagonal  Cook's D
4  -0.029151  -0.026991  0.605906  0.000653
```

Ve druhém případě jsou jackknife residuum i Cookova vzdálenost extrémně velké.

Row	Residual	Rstudent	Diagonal	Cook's D
4	-2.382046	-5.067308	0.605906	4.361897

Tento bod tedy má zásadní vliv na hodnoty odhadů, což je ostatně na první pohled vidět v této úloze, kdy model má je jeden regresor, i na grafech proložených regresních přímek. V případě modelů s více regresory ovšem takové vizuální posouzení není možné a regresní diagnostika je užitečným nástrojem pro ověření předpokladů a posouzení vhodnosti modelu.

Důsledky umístění bodu 4 vidíme v následující tabulce směrodatných odchylek:

	1 (vlivný)	2 (odlehlý)	bez bodu 4
intercept	0.849	1.951	1.152
směrnice	0.130	0.299	0.211
sm. odch. reziduí	1.15	2.65	1.25

Shrnutí

- *projekční matice, ortogonální projekce*
- *rozklad celkového součtu čtverců, index determinace* R^2
- *residua, jackknife residua, diagonální prvky projekční matice, Cookova vzdálenost*
- *autokorelace, Durbin–Watsonova statistika*

Kontrolní otázky

1. *Zkuste dokázat, že opravdu platí* $TSS = MSS + RSS$.
2. *Načrtněte, co znamená ortogonalita vektorů* $\hat{\mathbf{y}}$ *a* $\mathbf{e}$. *Pro graf zvolte model se dvěma regresory a výběr o rozsahu 3.*
3. *Jaká hypotéza se testuje v analýze rozptylu, uváděné ve výstupu statistických programů pro lineární regresi?*
4. *K čemu slouží regresní diagnostika?*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.