Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Jana Burešová

Regresní analýza nákladovosti správních úřadů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Milan Vítek, Finanční úřad Praha - západ

Studijní program: Matematika

Studijní obor: Finanční matematika

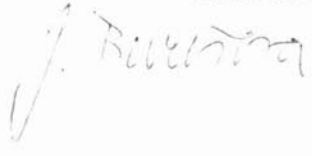
2006

Děkuji panu RNDr. Milanu Vítkovi za odborné vedení mé bakalářské práce.

Prohlašuji, že jsem svou bakalářskou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 28.5.2006

Jana Burešová



Obsah

Název práce: Regresní analýza nákladovosti správních úřadů

Autor: Jana Burešová

Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Milan Vítek, Finanční úřad Praha - západ

e-mail vedoucího: milan.vitek@prz.pr.ds.mfcr.cz

Abstrakt: Regresní analýza zpracovaná statistickým softwarem odhalila, že celkové náklady finančních úřadu závisí především na počtu zaměstnanců (náhodná veličina), dále pak na výši nájemného (nenáhodná veličina) a nepatrně i na poloze úřadu vzhledem k oblasti. Proměnné jako počet daňových subjektů, počet odeslaných písemností, velikost oblasti atd. byly vyloučeny buď z důvodu zabránění multikolinearitě nebo z důvodu nedůležitosti pro normální lineární model (tj. nebyla zamítnuta hypotéza o nulovosti dílčího regresního koeficientu). Za účelem hospodárnosti by tedy počty zaměstnanců finančních úřadů měly co nejvíce odpovídat počtu příslušných daňových subjektů, který s počtem zaměstnanců silně koreluje.

Klíčová slova: regrese, multikolinearita, koeficient determinace, správní úřad

Title: Regression Analysis of Administrative Agencies´ Costs

Author: Jana Burešová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Milan Vítek, Finanční úřad Praha - západ

Supervisor's e-mail address: milan.vitek@prz.pr.ds.mfcr.cz

Abstract: Multiple regression analysis provided by the statistical software NCSS revealed that the administrative agencies´ costs depend on the number of employees at most (variable costs), then also on the rent (fixed costs) and a bit on the position of the agency in the supervised area. Some of the other factors (number of tax subjects, number of letters, area, etc.) were excluded from the model because of the multicollinearity or the non-rejection of the hypothesis that the partial regression coefficient equals zero. Due to the hard correlation between the number of employees and the number of tax subjects these two numbers should correspond as most as possible to fulfil the principle of efficiency.

Keywords: regression, multicollinearity, coefficient of determination, administrative agency

1 Úvod

Daňový systém je jednou z oblastí, která významným způsobem ovlivňuje rozhodování ekonomických subjektů i mínění veřejnosti. Při tvorbě a řízení daňové politiky je proto nutno klást důraz na efektivnost. Efektivní daňový systém by měl fungovat tak, aby na jedné straně přinášel prostředky do veřejných rozpočtů a na straně druhé co nejméně zatěžoval subjekty, které prostředky odvádějí, ale i subjekty, které tyto prostředky spravují, tedy správní úřady.

V České republice je otázka efektivnosti umocněna nevyrovnaností státního rozpočtu, a proto jsou ve snaze minimalizovat rozpočtové výdaje prováděny četné rozbory a analýzy hospodaření správních úřadů.

V bakalářské práci bude za použití regresní analýzy vypracována statistická analýza závislosti výše nákladů správních úřadů na jiných faktorech. Jako statistické jednotky budou sloužit finanční úřady Středočeského kraje, kterých v roce, z něhož jsou čerpána data, tj. v roce 2005 bylo třicet.

Porovnáním celkových nákladů jednotlivých úřadů s počtem pracovníků, počtem daňových subjektů a dalšími ukazateli bude zkoumáno, zda hospodárnost, šetrnost a obezřetnost jsou skutečně zásadami uplatňovanými u státních výdajů ve formě výdajů na daňovou administrativu.

2 Statistický model

2.1 Termín regresní analýza

Pojem regrese (krok zpět, zpětný postup) se používá ve statistice ke zkoumání vztahu mezi jednou náhodnou veličinou na straně jedné (odezva, závisle proměnná) a jednou nebo více náhodnými veličinami na straně druhé (regresory, prediktory, nezávisle proměnné), jinak řečeno závislost vysvětlované proměnné na proměnných vysvětlujících.

Úkolem regresní analýzy je najít „idealizující" matematickou funkci tak, aby co nejlépe vyjadřovala charakter závislosti. Tato matematická funkce se nazývá regresní funkce. Podobně jako anglický přírodovědec Francis Galton (1822-1911) při vyšetřování závislosti tělesné výšky synů na tělesné výšce otců se budeme zabývat podmíněným rozdělením náhodné veličiny Y při pevné hodnotě x náhodné veličiny X, regresní funkcí se pak stává střední hodnota $E(Y|X = x)$. Současně se zkoumáním průběhu závislosti nás zajímá i síla (intenzita) závislosti, která rovněž vypovídá o kvalitě regresní funkce.

2.2 Lineární model

Nejčastěji využívaným modelem regresní analýzy bývá lineární model (model lineární v parametrech), jehož maticový zápis je

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

kde

- $\mathbf{Y} = (y_1,\ldots,y_n)'$ představuje n-složkový náhodný vektor pozorování, tj. hodnoty vysvětlované proměnné.
- $\mathbf{X} = (x_{ij})_{j=1,\ldots,k}^{i=1,\ldots,n}$ je matice známých hodnot vysvětlujících proměnných (nebo jejich funkcí), jejichž počet je roven k, $k < n$. Předpokládáme, že hodnost matice je rovna k, tedy uvažujeme model s úplnou hodností.

- $\boldsymbol{\beta} = (\beta_1,...,\beta_k)'$ je k-rozměrný vektor, jehož jednotlivé složky jsou nenáhodné veličiny označované jako regresní koeficienty. Tyto budeme odhadovat.
- $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_n)'$ označuje náhodný vektor chyb, pro nějž platí $\mathrm{E}\boldsymbol{\varepsilon} = (0,...,0)$ a $\mathrm{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$, kde $\mathbf{I}$ je jednotková matice typu $n \times n$.

Popsaný model lze souhrnně zapsat jako

$$\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

neboť $\mathrm{E}\mathbf{Y} = \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathrm{E}\boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}$ a $\mathrm{var}\,\mathbf{Y} = \mathrm{var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathrm{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$.

Nejjednodušším lineárním modelem (po konstantní závislosti) je regresní přímka $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1,...,n$. Pak tedy $k = 2$, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}$,

$\mathbf{Y} = (y_1,...,y_n)'$ a $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_n)'$. Vektor jedniček je součástí matice téměř vždy.

Některé nelineární modely se dají na lineární převést jednoduchou transformací. Příklady jsou uvedené v následující tabulce:

tabulka 1 Transformace nelineární regresní funkce

Funkce	Linearizující transformace
$y = \beta_0 x^{\beta_1}$	$\ln y = \ln \beta_0 + \beta_1 \ln x$
$y = \beta_0 e^{\beta_1 x}$	$\ln y = \ln \beta_0 + \beta_1 x$
$y = \dfrac{x}{\beta_0 + \beta_1 x}$	$\dfrac{x}{y} = \beta_0 + \beta_1 x$

Odhad regresních koeficientů

Jak bylo uvedeno, vektor regresních koeficientů je neznámý, ale lze jej odhadnout. Pro odhad těchto koeficientů a zároveň i regresní funkce se nejčastěji

používá *metoda nejmenších čtverců*, která se snaží co nejvíce přiblížit pozorované hodnoty k hodnotám vyrovnaným (hodnotám stanoveným modelem).

Tato metoda má za cíl minimalizovat výraz $\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2$, maticovým zápisem $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ pro všechny hodnoty parametru $\boldsymbol{\beta}$. Spočteme první derivaci výrazu podle $\boldsymbol{\beta}$, položíme ji rovnu nule a získáme tak odhad, který budeme značit symbolem $\hat{\boldsymbol{\beta}}$.

$$\frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) =$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})'\mathbf{Y} + (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}\right] =$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}\right] =$$

$$= -2\mathbf{Y}'\mathbf{X} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{Y}'\mathbf{X} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

Tato maticová rovnice se nazývá *normální rovnice* (soustava normálních rovnic). Jelikož uvažujeme model s úplnou hodností, tj. hodnost matice $\mathbf{X}$ rovnu k, je matice $\mathbf{X}'\mathbf{X}$ regulární. Tu můžeme invertovat a získáme tak jediný odhad vektoru $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. V modelu s neúplnou hodností by existovalo více řešení normální rovnice, tedy více odhadů vektoru $\boldsymbol{\beta}$.

Spočteme-li i druhou derivaci výrazu, získáme positivně definitní matici $2\mathbf{X}'\mathbf{X}$, čímž jsme se přesvědčili, že nalezený odhad je skutečně minimem.

Jaké jsou vlastnosti tohoto odhadu? $\hat{\boldsymbol{\beta}}$ je nestranným odhadem vektoru $\boldsymbol{\beta}$, neboť $\mathrm{E}\hat{\boldsymbol{\beta}} = \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$.

$$\mathrm{var}\hat{\boldsymbol{\beta}} = \mathrm{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{var}\,\mathbf{Y}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} =$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Další použité veličiny

Pojďme se dále zabývat vektorem $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, tedy vektorem vyrovnaných hodnot. Označme $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, pak můžeme zapisovat $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. Matice $\mathbf{H}$ je symetrická a idempotentní, tj. $\mathbf{H}' = \mathbf{H}$ a $\mathbf{H}\mathbf{H} = \mathbf{H}$, neboť

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

a $\mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$.

Spočtěme si rovněž střední hodnotu a rozptyl vektoru vyrovnaných hodnot.

$\mathrm{E}\hat{\mathbf{Y}} = \mathrm{E}\mathbf{H}\mathbf{Y} = \mathbf{H}\mathrm{E}\mathbf{Y} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = \mathrm{E}\mathbf{Y}$, tj. $\hat{\mathbf{Y}}$ je nestranným odhadem vektoru středních hodnot $\mathrm{E}\mathbf{Y}$.

$\mathrm{var}\,\hat{\mathbf{Y}} = \mathrm{var}\,\mathbf{H}\mathbf{Y} = \mathbf{H}\,\mathrm{var}\,\mathbf{Y}\mathbf{H} = \sigma^2\mathbf{H}\mathbf{H} = \sigma^2\mathbf{H}$

Označme $\mathbf{u}$ vektor residuí, tedy $\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. Všimněme si, že matice $\mathbf{I} - \mathbf{H}$ je rovněž symetrická a idempotentní, neboť platí

$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - 2\mathbf{H} + \mathbf{H}\mathbf{H} = \mathbf{I} - 2\mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}$.

Pro určení kvality modelu je potřebná náhodná veličina SS_e, která se nazývá residuální součet čtverců. Definujme ji vztahem $SS_e = \mathbf{u}'\mathbf{u}$, platí pro ni tedy i rovnosti $SS_e = \mathbf{Y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x_{ij}\hat{\beta}_j)^2$. Obecně platí, čím menší residuální součet čtverců, tím lepší model.

2.3 Normální lineární model

Dále budeme předpokládat, že $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, tedy že používáme *normální lineární model*. Tento předpoklad nám pomůže určit rozdělení dalších veličin. Konkrétně

- $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$,
- $\mathbf{u} \sim N(0, \sigma^2(\mathbf{I} - \mathbf{H}))$,
- $\hat{\mathbf{Y}} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H})$,

- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $SS_e / \sigma^2 \sim \chi^2_{n-k}$.

První čtyři rozdělení jsou zřejmá, dokážeme tedy pouze poslední rozdělení pomocí tvrzení: pokud je $\mathbf{A}\text{var}\mathbf{Z}$ idempotentní, pak $\mathbf{Z}'\mathbf{A}\mathbf{Z} \sim \chi^2_{tr(\mathbf{A}\text{var}\mathbf{Z})}$ (bez důkazu). Hodnota $tr(A)$ znamená stopa matice A, tj. součet prvků na diagonále matice.

$SS_e = \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y} = (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{H})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$, neboť

$(\mathbf{I}-\mathbf{H})\mathbf{X} = \mathbf{X}-\mathbf{H}\mathbf{X} = \mathbf{X}-\mathbf{X} = \mathbf{0}$.

Položíme-li $\mathbf{Z} = \mathbf{Y}-\mathbf{X}\boldsymbol{\beta}$ a $\mathbf{A} = (\mathbf{I}-\mathbf{H})/\sigma^2$, získáme $\mathbf{Z}'\mathbf{A}\mathbf{Z} = SS_e/\sigma^2$. Variační matice $\text{var}\,\mathbf{Z} = \sigma^2\mathbf{I}$, neboť $\mathbf{Z} = \boldsymbol{\varepsilon}$. Matice $\mathbf{A}\,\text{var}\,\mathbf{Z} = \mathbf{I}-\mathbf{H}$ je idempotentní, jak jsme již výše ukazovali. Zbývá tedy ještě dokázat, že počet stupňů volnosti je skutečně roven $n-k$. $tr(\mathbf{A}\,\text{var}\,\mathbf{Z}) = tr(\mathbf{I}-\mathbf{H}) = tr(\mathbf{I}) - tr(\mathbf{H}) = n-k$, jelikož pro každou idempotentní matici, tedy i matici $\mathbf{H}$, platí, že stopa matice je rovna její hodnosti.

Z vlastností χ^2-rozdělení víme, že jeho střední hodnota se rovná počtu stupňů volnosti, v našem případě $\mathrm{E}\frac{\mathrm{SS_e}}{\sigma^2} = n-k$. Pak rovněž platí $\mathrm{E}\frac{\mathrm{SS_e}}{n-k} = \sigma^2$, tedy $\hat{\sigma}^2 = \frac{\mathrm{SS_e}}{n-k}$ je nestranným odhadem σ^2.

2.4 Testování hypotéz

Testování hypotéz o vektoru parametrů $\boldsymbol{\beta}$ se používá při odhadování typu závislosti. Pokud si například nejsme jisti, zda je závislost náhodné veličiny Y na náhodné veličině X kvadratická $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$, nebo pouze lineární $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, zjednodušeně nás zajímá, zda je $\beta_3 = 0$ či ne. Pak použijeme t-test, který vyplývá z následujícího tvrzení.

Nechť $\mathbf{c} \in \mathrm{R}^k$. Pak $\dfrac{\mathbf{c}'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sqrt{\dfrac{SS_e}{n-k}\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-k}$.

Tuto veličinu se Studentovým t-rozdělením o n-k stupních volnosti získáme jako podíl dvou nezávislých veličin, z nichž jedna má normované normální rozdělení

a druhá χ^2-rozdělení o n-k stupních volnosti (vydělena počtem stupňů volnosti a odmocněna). Veličinu v čitateli získáme touto úvahou. Jelikož $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, pak $\mathbf{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})$ a $\frac{\mathbf{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{\sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim N(0,1)$. Veličina ve jmenovateli je již známa, $SS_e / \sigma^2 \sim \chi^2_{n-k}$. Jejich úpravou a podělením získáme výše uvedený zlomek.

$$\frac{\frac{\mathbf{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{\sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}}}{\sqrt{\frac{SS_e / \sigma^2}{n-k}}} = \frac{\mathbf{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{\frac{SS_e}{n-k}\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}}$$

Vlastní t-test má nulovou hypotézu $\mathbf{c}'\boldsymbol{\beta} = \delta$ proti alternativní hypotéze $\mathbf{c}'\boldsymbol{\beta} \neq \delta$ a zkoumá testovou statistiku $T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \delta}{\sqrt{\frac{SS_e}{n-k}\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}}$, která má za platnosti nulové hypotézy t-rozdělení o n-k stupních volnosti. Na hladině spolehlivosti α zamítáme nulovou hypotézu, právě když $|T| \geq t_{n-k}(1-\frac{\alpha}{2})$, kde $t_{n-k}(1-\frac{\alpha}{2})$ značí $(1-\frac{\alpha}{2})$*100%-ní kvantit Studentova t-rozdělení o n-k stupních volnosti. Vrátíme-li se zpět k příkladu, pak testujeme platnost nulové hypotézy pro $\mathbf{c} = (0,0,1)'$ a $\delta = 0$.

Při interpretaci nulové hypotézy $\beta_i = 0$ je třeba vzít v úvahu, že všechny ostatní vysvětlující proměnné v modelu zůstávají zachovány. Nejde tedy o testování prostého tvrzení, že na i-té veličině střední hodnota náhodné veličiny Y nezávisí. [2] Jde spíš o uvážení, zda je možné i-tou veličinu z modelu vyloučit.

2.5 Koeficient determinace

O kvalitě použitého modelu vypovídá *koeficient determinace* R^2, který je definován vztahem $R^2 = 1 - \frac{SS_e}{\sum_{i=1}^{n}(y_i - \bar{Y})^2}$, kde $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Suma ve jmenovateli představuje celkový součet čtverců, který získáme součtem reziduálního a teoretického (modelového) součtu čtverců. Pak můžeme koeficient determinace

interpretovat jako poměr mezi teoretickým a celkovým součtem čtverců. Tento koeficient ukazuje, jakou část variability závisle proměnné se pomocí uvažované závislosti podařilo vysvětlit variabilitou nezávisle proměnných. Proto se často hodnota koeficientu determinace udává v procentech.

Při hodnocení velikosti koeficientu determinace je užitečné znát jeho rozdělení v případě, že nekonstantní regresory v $\mathbf{X}$ nepřispívají k vysvětlení variability vektoru $\mathbf{Y}$. [1] Předpokládejme model $\mathbf{Y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$ splňující $k > 1$ a $\mathbf{X}$ obsahuje konstantní sloupec. Potom má koeficient determinace R^2 beta rozdělení s parametry $(k-1)/2, (n-k)/2$ a náhodná veličina $F = \frac{R^2}{1-R^2}\frac{n-k}{k-1}$ má rozdělení $F_{k-1,n-k}$.

2.6 Multikolinearita

Termín multikolinearita (kolinearita) pochází z ekonomických aplikací, i když se dnes s tímto problémem setkáváme například také v technických aplikacích. Zhruba řečeno, multikolinearitou se míní případ, kdy sloupce matice $\mathbf{X}$ (jednotlivé regresory) jsou téměř lineárně závislé. K tomu může dojít, když nepřesně zjistíme či zapíšeme (pomocí konečného počtu číslic) prvky matice $\mathbf{X}$. Častěji jde o přirozenou vlastnost vyšetřovaných jevů.

Na multikolinearitu mohou upozornit velké hodnoty dílčích korelačních koeficientů mezi regresory. Častým projevem multikolinearity jsou malé hodnoty testové statistiky T pro jednotlivá β_i, přestože koeficient determinace je významně nenulový. [1]

Z věcného hlediska to znamená, že velké množství regresorů v modelu či jejich přidávání nemusí být účelné. Takový model může být obtížně interpretovatelný a znehodnocený multikolinearitou. Dá se říci, že vysvětlující proměnná, která silně koreluje s jinou, víceméně jen opakuje informaci, která je již v modelu obsažena, ale zato rychle snižuje naději modelu na kvalitní odhad parametrů.

3 Statistické jednotky

3.1 Veřejná správa

Pojem správní úřad zahrnuje množství orgánů, které jsou činné ve veřejné správě, tj. ve státní správě, samosprávě a jiné veřejné správě. Státní správa je vykonávána jménem státu, v jeho zájmu a disponuje prostředky státně mocenské povahy. Orgány státu vykonávající státní správu (tj. správní úřady v užším smyslu) můžeme dělit podle způsobu rozhodování, místní působnosti, rozsahu věcné působnosti a dalších kritérií. Podle příslušných právních předpisů k nim řadíme především

- vládu (správní orgán se všeobecnou působností rozhodující ve sboru),
- ústřední správní orgány (ministerstva a další ústřední správní orgány jako např. Český statistický úřad či Národní bezpečnostní úřad),
- další správní úřady s celostátní působností (podřízené příslušnému ministerstvu, např. Česká inspekce životního prostředí, Česká správa sociálního zabezpečení, Národní archiv),
- místní správní úřady jako
    - specializované územní správní úřady,
    - krajské úřady a další orgány krajů, pokud vykonávají státní správu,
    - obecní úřady a další orgány obcí, pokud vykonávají státní správu.

Specializované územní správní úřady představují širokou škálu správních úřadů působících na území okresu, kraje nebo zákonem jinak vymezeného prostoru (př. vojenské újezdy). Tyto správní úřady charakterizujeme jako monokratické orgány s dílčí specializovanou věcnou působností a omezenou územní působností. Zmíněné úřady jsou zpravidla podřízeny příslušnému ministerstvu. Zřízeny jsou vždy zákonem v souladu s článkem 79 Ústavy České republiky. Příkladem jsou finanční úřady, celní úřady, katastrální úřady či úřady práce.

3.2 Popis statistických jednotek

Jak již bylo řečeno, pro účely bakalářské práce budou za statistické jednotky sloužit finanční úřady ve Středočeském kraji, tj. finanční úřady v působnosti Finančního ředitelství v Praze, kterých bylo v roce 2005 třicet. Podle §1 zákona č.531/1990 Sb., o územních finančních orgánech, ve znění pozdějších předpisů, jsou územní finanční orgány (tj. finanční úřady a finanční ředitelství) správními úřady, které:

- vykonávají správu daní a správních poplatků jimi vyměřovaných a vybíraných,
- spravují dotace,
- provádějí finanční revize,
- provádějí cenovou kontrolu podle zvláštního právního předpisu,
- provádějí řízení o přestupcích v oboru své působnosti,
- vybírají a vymáhají odvody, poplatky, úhrady, úplaty, pokuty a penále, včetně nákladů řízení, které jsou uloženy jinými orgány státní správy, s výjimkou pokut ukládaných obcemi a kraji, apod.

3.3 Výběr dat

Některé z výše uvedených zákonem předepsaných činností jsou spojeny s vyššími náklady. Jedná se především o mzdové náklady (včetně zdravotního a sociálního pojištění), provozní náklady (nájemné), náklady na provádění finanční revize (cestovné) či náklady na nákup jiných služeb (poštovné). Některé náklady jsou fixní (nájem), jiné variabilní, tudíž závislé na počtu zaměstnanců či daňových subjektů. Náklady jako například cestovné se jistě odvíjí od velikosti oblasti, kterou úřad spravuje, popřípadě na poloze úřadu vzhledem k této oblasti. Rozdílné náklady spojené s rozdílnou dostupností míst a různou hustotou silniční a železniční sítě se zdají být ve Středočeském kraji zanedbatelné.

Z účetních podkladů za kalendářní rok 2005 bychom se měli dále zabývat následujícími osmi položkami:

- Celková výše nákladů [v Kč] – součet mzdových nákladů, nákladů na zdravotní a sociální pojištění a provozních (neinvestičních) nákladů.
- Počet daňových subjektů (dále jen DS) příslušných jednotlivým úřadům – dle §6 zákona č.337/1992 Sb., o správě daní a poplatků, ve znění pozdějších předpisů, se daňovým subjektem předně rozumí poplatník (osoba, jejíž příjmy, majetek nebo úkony jsou přímo podrobeny dani) a plátce daně (osoba, která pod vlastní majetkovou odpovědností odvádí správci daně daň vybranou od poplatníků). Pro jednoduchost předpokládáme, že tento počet je v průběhu roku neměnný. Rovněž nebudeme rozdělovat subjekty podle jednotlivých daní.
- Počet zaměstnanců (označme PS-pracovní síla) – jak již bylo řečeno, mzdové náklady představují nedílnou složku celkových nákladů, dokonce největší složku pro každý úřad. Pro jednoduchost nerozlišujeme platové třídy a předpokládáme, že tento počet je v průběhu roku neměnný.
- Počet odeslaných písemností [v ks] – úřady rozesílají různé druhy písemností jako především doporučená psaní do vlastních rukou, jelikož den doručení je obvykle rozhodný pro počátek běhu lhůty, jejíž nesplnění by pro příjemce mohlo být spojeno s právní újmou. Podle zákona správce daně doručuje úřední písemnosti zpravidla poštou, a tak mu vznikají náklady v podobě poštovného.
- Počet odeslaných složenek [v ks] – platba daně z nemovitosti jako jediná probíhá prostřednictvím složenek. Pokud u poplatníka této daně nedošlo ke změně, nemusí podávat nové daňové přiznání, ale rovnou obdrží složenku k zaplacení. Jelikož poplatek za odeslání složenky je nižší než poplatek za doporučený dopis a jiné písemnosti, budeme tyto dvě skupiny rozlišovat.
- Počet provedených daňových kontrol (dále jen DK) [v ks] - daňovou kontrolou pracovník správce daně zjišťuje nebo prověřuje daňový základ nebo jiné okolnosti rozhodné pro správné stanovení daně. Činí tak přímo u daňového subjektu nebo na místě, kde je to vzhledem k účelu kontroly nejvýhodnější. DK se provádí v rozsahu nezbytně nutném pro dosažení účelu podle zákona o správě daní a poplatků.

- Počet provedených místních šetřeních (dále jen MS) [v ks] - v souvislosti s daňovým řízením může správce daně provádět MS jak u subjektu daně, tak i jiných osob. Pracovník správce daně má především právo na přístup do každé provozní místnosti a přístup k účetním písemnostem či jiným záznamům. MS má zpravidla kratší charakter než DK. Zatímco u DK se jedná o dny, u MS jsou to většinou hodiny, proto tyto dvě kontrolní činnosti finančních úřadů odlišíme.

- Výše nájmu [v Kč] – některé úřady nevlastní budovy, v nichž sídlí, a tak musí platit nájemné. Tyto náklady jsou u jednotlivých úřadů velice odlišné a navíc jako nutná náležitost nájemní smlouvy nepředstavují náhodnou veličinu. Nemůžeme je tedy zahrnovat mezi regresory. Odečteme je od celkových nákladů a tento rozdíl pak budeme považovat za vysvětlovanou proměnnou.

Do analýzy budou zahrnuty i údaje o velikost oblasti, kterou úřad spravuje, a poloze úřadu vzhledem k této oblasti. Rozloha oblastí vyplývá z údajů Českého statistického úřadu a z již citovaného zákona o územních finančních orgánech, údaje jsou uváděny v km^2. Polohu úřadu rozdělíme do tří kategorií, nejlepší poloha co se týče velikosti nákladů je poloha *přibližně ve středu oblasti* – hodnota 0, některé úřady ale leží spíše *na kraji oblasti* – hodnota 1, a některé dokonce *mimo oblast* – hodnota 2.

tabulka 2 – Data vybraná pro regresní analýzu

PS	DS	NAJEMNE	SLOZENKY	PISEMNOSTI	DK	MS	VYMERA	POLOHA	CELK_NAKLADY
21	13 983	0	4 500	5 546	322	87	208	0	7 610 187
21	16 252	360 948	9 100	7 733	136	52	289	0	8 080 412
22	20 440	0	7 900	10 517	163	104	137	0	8 445 985
25	17 292	202 376	8 000	10 771	368	94	173	1	9 144 418
25	20 045	0	7 480	12 431	352	153	449	0	10 039 173
27	21 744	1 306 846	9 400	12 372	227	145	318	0	11 364 125
28	24 429	1 717 478	10 900	13 351	148	126	283	1	11 176 888
29	19 171	50 150	4 000	15 384	248	208	233	0	10 547 690
30	21 850	0	6 500	11 958	158	249	274	0	11 439 455
30	24 774	0	8 500	12 433	592	213	302	1	11 385 107
36	21 327	0	5 400	19 046	454	214	140	1	13 650 130
38	26 264	1 196 281	9 074	18 167	386	150	123	0	14 088 658
42	26 095	0	10 400	16 197	297	255	536	1	14 385 205
42	26 383	1 724 937	9 050	22 725	168	295	174	0	16 975 528
43	28 669	0	11 500	19 116	316	306	296	0	15 273 325
47	29 915	0	9 000	17 588	316	286	349	1	17 310 987
49	35 549	0	3 500	24 394	515	310	369	0	20 471 205
50	44 201	0	12 000	23 681	356	176	340	1	20 705 369
51	35 949	821 988	18 300	22 625	432	312	227	1	19 555 085
51	57 931	1 981 636	5 000	28 894	284	286	377	1	19 281 158
65	57 781	0	21 971	36 797	270	257	366	0	24 026 942
68	57 609	0	24 000	43 137	241	338	455	2	26 300 808
68	61 989	1 939 650	17 000	35 695	552	319	406	0	25 703 529
69	67 567	0	23 665	36 595	329	276	649	0	23 487 475
80	67 519	0	22 700	39 656	485	628	896	0	27 709 002
83	46 994	0	17 700	34 954	521	436	417	0	29 887 143
89	87 330	4 243 878	11 500	42 001	812	558	440	0	31 564 053
96	74 720	0	22 100	45 421	930	583	596	0	34 289 692
121	100 830	0	31 190	67 544	601	1 181	351	0	43 567 953
151	124 373	0	59 732	85 121	399	244	581	2	55 206 975

4 Statistická analýza

4.1 Domněnky

Statistická šetření jsou často koncipována za účelem vyvrácení nebo potvrzení nějakých domněnek. My se před vlastním rozborem také zamyslíme nad tím, k jakým závěrům bychom asi měli dospět. Celkové náklady by měli nejvíce záviset na počtu zaměstnanců a ten by měl být za účelem efektivnosti co nejvíce závislý na počtu DS. Počet odeslaných písemností a složenek by měl souviset s počtem DS, tak bychom se měli střežit multikolinearity. Zároveň budeme pozorovat hodnoty T-statistiky pro jednotlivé regresory a tím odhadovat jejich nezbytnou přítomnost v modelu.

4.2 Závislost mezi regresory

V kapitole 2.6 jsme hovořili o nebezpečnosti velkého množství regresorů. V našem případě je osm relativně dost, a proto bychom se měli ve snaze předejít multikolinearitě podívat, jak jsou jednotlivé regresory lineárně (ne)závislé, tj. jaké jsou jejich korelační koeficienty.

tabulka 3 - Tabulka korelačních koeficientů

	DK	DS	MS	PISEMNOSTI	POLOHA	PS	SLOZENKY	VYMERA
DK	**1,0000**	0,5284	0,5976	0,4661	-0,1180	0,5554	0,2399	0,3620
DS	0,5284	**1,0000**	0,6653	0,9393	0,1818	0,9656	0,8392	0,6160
MS	0,5976	0,6653	**1,0000**	0,6638	-0,1777	0,6965	0,4105	0,4215
PISEMNOSTI	0,4661	0,9393	0,6638	**1,0000**	0,2393	0,9821	0,8935	0,5554
POLOHA	-0,1180	0,1818	-0,1777	0,2393	**1,0000**	0,1869	0,3563	0,0448
PS	0,5554	0,9656	0,6965	0,9821	0,1869	**1,0000**	0,8732	0,5928
SLOZENKY	0,2399	0,8392	0,4105	0,8935	0,3563	0,8732	**1,0000**	0,5340
VYMERA	0,3620	0,6160	0,4215	0,5554	0,0448	0,5928	0,5340	**1,0000**

Tato symetrická tabulka ukazuje silnou závislost ve skupině proměnných DS-PISEMNOSTI-PS-SLOZENKY (potvrzení domněnky ze 4.1). Z této skupiny tak stačí vybrat jednoho zástupce, který v regresní analýze bude vypovídat za všechny

čtyři. Zvolíme veličinu PS, která má největší korelační koeficient s vysvětlovanou proměnnou. Cor(PS;CELK_NAKLADY - NAJEMNE) = 0,993

4.3 Model

K této analýze budeme používat normální lineární model (tak, jak je popsaný v kapitole 2.3). Po vyloučení tří proměnných v kap. 4.2 a zahrnutí sloupce jedniček jako absolutních členů je $k = 6$, $n = 30$. Vektor $\mathbf{Y}$ představuje rozdíl mezi hodnotami CELK_NAKLADY a NAJEMNE z výše uvedené tabulky (označme jej VYSVETLOVANA), matice $\mathbf{X}$ (typu 30x6) je složená ze sloupečků jedniček, PS, DK, MS, VYMERA a POLOHA. $\boldsymbol{\beta} = (\beta_1,...,\beta_6)$, $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_{30})$.

4.4 Vyhodnocení statistickým softwarem

K výpočtu odhadů regresních koeficientů, koeficientu determinace, k testování hypotéz o vektoru parametrů a dalším výpočtům budeme používat statistický software NCSS, který je k dispozici v počítačové laboratoři na fakultě.

Po zadání veškerých údajů do tohoto softwaru a nastavení regresní analýzy poskytl NCSS tyto výsledky:

a) Šest regresorů

Run Summary Section

Parameter	Value
Dependent Variable	VYSVETLOVANA
Number Ind. Variables	5
Weight Variable	None
R2	0,9881
Adj R2	0,9856
Coefficient of Variation	0,0691
Mean Square Error	1,767369E+12
Square Root of MSE	1329424
Ave Abs Pct Error	4,545

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5,0%?	Power of Test at 5,0%
Intercept	841218,9055	695001,5132	1,210	0,2379	No	0,2134
DK	-1547,3340	1702,3202	-0,909	0,3724	No	0,1408
MS	-88,2122	1818,6730	-0,049	0,9617	No	0,0502
POLOHA	542674,4297	452017,0562	1,201	0,2416	No	0,2107
PS	359368,2862	14134,0200	25,426	0,0000	Yes	1,0000
VYMERA	-994,0062	1804,3684	-0,551	0,5868	No	0,0826

Z první části si všimněme hlavně řádku R2, který udává hodnotu koeficientu determinace R^2. Můžeme tedy říci, že 98,81% variability vysvětlované proměnné se modelem podařilo vysvětlit. Jelikož ale dále budeme měnit počet vysvětlujících proměnných, ke srovnávání modelů musíme použít hodnotu Adj R2. Tzv. očištěný koeficient determinace $R^2_{Adjusted} = 1-(1-R^2)\frac{n-1}{n-k-1}$ je různému počtu veličin přizpůsobený. Zde Adj R2 = 98,56%.

Z druhé části nás v této fázi zajímá hlavně čtvrtý sloupec, který udává velikost T-statistiky z testování hypotézy o nulovosti dílčího regresního koeficientu. Jediný regresor, pro který je tato hypotéza zamítnuta je PS, což svědčí o tom, že PS bude nejdůležitějším článkem výsledného modelu. Nejnižší hodnota T-statistiky je u MS - tato proměnná je pro model nejméně přínosná.

b) Pět regresorů

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5,0%?	Power of Test at 5,0%
Intercept	840146,9884	680648,6357	1,234	0,2286	No	0,2207
DK	-1569,7894	1605,1386	-0,978	0,3375	No	0,1559
POLOHA	550999,4464	409734,2838	1,345	0,1908	No	0,2531
PS	358972,9973	11315,1072	31,725	0,0000	Yes	1,0000
VYMERA	-991,1402	1767,0512	-0,561	0,5799	No	0,0840

Adj R2 = 98,62% - hodnota vzrostla, model po vyloučení MS je kvalitnější. Vyloučíme další veličinu s nejmenší T-statistikou, tedy veličinu VYMERA.

c) Čtyři regresory

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5,0%?	Power of Test at 5,0%
Intercept	656766,4556	589065,5760	1,115	0,2751	No	0,1890
DK	-1594,4156	1583,2480	-1,007	0,3232	No	0,1628
POLOHA	567752,5105	403222,0521	1,408	0,1710	No	0,2735
PS	355782,7039	9651,8538	36,862	0,0000	Yes	1,0000

Adj R2 = 98,6564% opět vzrostl, z modelu vyřadíme DK.

d) Tři regresory

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5,0%?	Power of Test at 5,0%
Intercept	310566,7909	478473,7145	0,649	0,5218	No	0,0960
POLOHA	677984,4001	388182,3553	1,747	0,0921	No	0,3916
PS	350029,2380	7781,3705	44,983	0,0000	Yes	1,0000

Očištěný koeficient determinace nepatrně poklesl, Adj R2 = 98,6557%. Kdybychom odebrali i POLOHA (absolutní člen odebrat nemůžeme), tj. uvažovali bychom lineární závislost pouze na PS, koeficient by klesl daleko víc. Z posledních údajů můžeme také vyčíst, že hypotézu nulového regresního koeficientu pro proměnnou POLOHA můžeme zamítnout na hladině spolehlivosti $\alpha = 10\%$ (p-hodnota je rovna 0,0921).

4.5 Výsledný model

I když je očištěný koeficient determinace v bodě d) nepatrně horší než v bodě c), jako kvalitnější se jeví model z bodu d). V modelu c) je zarážející záporný regresní koeficient u proměnné DK. Provedené daňové kontroly mohou zvýšit výnosy, ale ne snížit náklady finančního úřadu.

Rovnice nejlepšího modelu je:

VYSVETLOVANA = 310567 + 677984*POLOHA + 350029*PS + ε

Koeficient determinace (neočištěný) je v tomto případě roven 98,75%. Získá se podílem teoretického (modelového) a celkového součtu čtverců (kap. 2.5) z tabulky 4. Podrobným prozkoumáním té samé tabulky zjistíme, že proměnná POLOHA vysvětluje velmi malou část modelu, pouze 0,14%. Z modelu jsme ji nevyloučili jen kvůli zachování vysokého koeficientu determinace. Nelze jí tedy klást příliš velký význam.

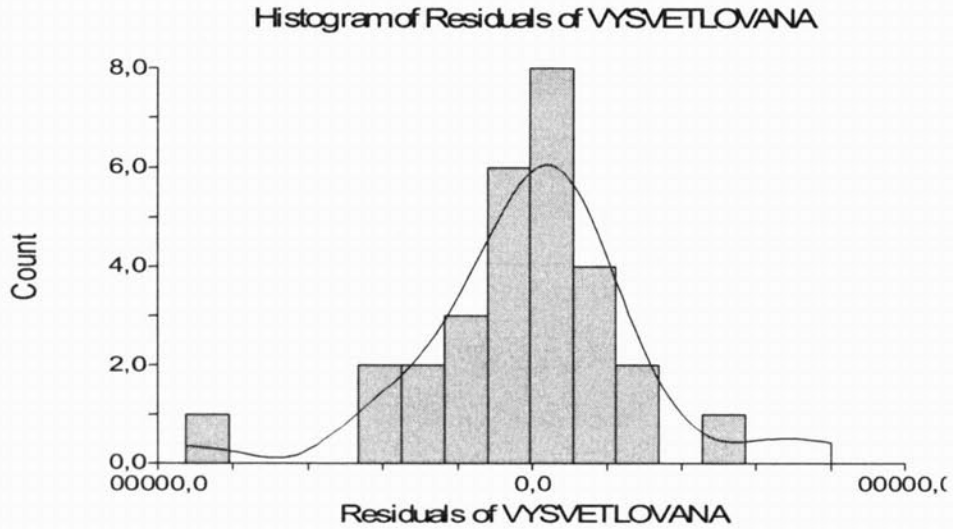
tabulka 4 - Analýza rozptylu výsledného modelu

Model Term	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5,0%)
Intercept	1		1,110254E+16	1,110254E+16			
Model	2	0,9875	3,521301E+15	1,760651E+15	1065,140	0,0000	1,0000
POLOHA	1	0,0014	5,042367E+12	5,042367E+12	3,050	0,0921	0,3916
PS	1	0,9380	3,344746E+15	3,344746E+15	2023,469	0,0000	1,0000
Error	27	0,0125	4,463036E+13	1,652976E+12			
Total	29	1,0000	3,565931E+15	1,229632E+14			

Měli bychom také ověřit některé předpoklady normálního lineárního modelu u modelu získaného. Předpoklad modelu s úplnou hodností je dodržen, hodnost matice **X** složené z vektorů jedniček, POLOHA a PS má skutečně hodnost rovnu třem.

Skutečnost, že vektoru reziduí má normální rozdělení se střední hodnotou nula, ověřuje graf na Obrázku 1, i když malý počet statistických jednotek způsobil jistou deformaci rozdělení. Tímto jsme ověřili i rozdělení $\mathbf{Y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$.

Obrázek 1 - Histogram reziduí získaného modelu



Jak je uvedeno v kapitole 2.3, díky předpokladu normálního rozdělení vysvětlované proměnné lze odhadnout i její rozptyl pomocí podílu $\hat{\sigma}^2 = \frac{SS_e}{n-k}$.

$$\hat{\sigma}^2 = \frac{SS_e}{n-k} = \frac{3{,}5213 * 10^{15}}{30-3} = 1{,}3042 * 10^{14}$$

Na tomto místě je vhodné potvrdit i volbu PS jako zástupce skupiny proměnných, které jsou navzájem závislé. Kdybychom místo PS zvolili DS z toho důvodu, že vlastně na DS závisí počet zaměstnanců i počet odeslaných složenek a písemností, tak by model vykazoval daleko nižší koeficient determinace (kolem 90%).

Další údaje, které nám statistické softwary nabízejí, jsou například 95%-ní intervaly spolehlivosti pro odhadnuté regresní koeficienty (tab.5), korelační matice odhadnutých koeficientů (tab.6) nebo výpis řádků, u nichž daným modelem vznikla neobvykle veliká rezidua (tab.7). Z tohoto výpisu lze vyčíst, že finanční úřad, jehož údaje se nacházejí na 17. řádku, má neobvykle vysoké náklady a v rámci efektivnosti systému by je měl snížit. Naopak úřad na 27. řádku má neobvykle nízké náklady. Účelem dalšího bádání by tedy mělo být zjišťování důvodů pro tyto velké odchylky od modelu, jestli došlo k nějakým mimořádným událostem (které by potvrdily

funkčnost modelu), nebo jestli je toto vybočení trvalé, pak jaké má příčiny a jak jim předcházet.

tabulka 5

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	310567,0	478474,0	-671182,0	1,29232E6
poloha	677984,0	388182,0	-118502,0	1,47447E6
ps	350029,0	7781,37	334063,0	365995,0

tabulka 6

Correlation matrix for coefficient estimates

	CONSTANT	poloha	ps
CONSTANT	1,0000	-0,1898	-0,8000
poloha	-0,1898	1,0000	-0,1869
ps	-0,8000	-0,1869	1,0000

tabulka 7

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
17	2,04712E7	1,7462E7	3,00921E6	2,66
27	2,73202E7	3,14632E7	-4,14299E6	-4,44

5 Závěr

Výši celkových nákladů třiceti finančních úřadů ve Středočeském kraji jsme původně chtěli vysvětlovat devíti faktory. Po uvážení nenáhodnosti nákladů ve formě nájemného jsme se rozhodli pro změnu vysvětlované proměnné na rozdíl mezi celkovými náklady a nájemným. Abychom předešli multikolinearitě, jevu, který zkresluje model, vyloučili jsme další tři vysvětlující proměnné, které silně korelovaly s počtem zaměstnanců. Konkrétně se jednalo o počty daňových subjektů, odeslaných písemností a složenek.

Testování hypotéz o nulovosti dílčích regresních koeficientů odhalilo, že proměnné počet místních šetření, rozloha podřízeného území a počet provedených daňových kontrol nejsou pro model přínosné. Výsledný model tedy kromě konstanty závisí jen na počtu zaměstnanců a nepatrně také na poloze úřadu vzhledem k oblasti.

VYSVETLOVANA = 310567 + 677984*POLOHA + 350029*PS + ε je rovnice modelu, pro který jsme se rozhodli. Koeficient determinace rovný 98,75% naznačuje, že tento model má vysokou vypovídací schopnost co se týče proměnlivosti vysvětlované proměnné. Tuto proměnlivost v daleko větší míře vysvětluje počet zaměstnanců než poloha úřadu.

Můžeme tedy závěrem říci, že výše celkových nákladů jednotlivých úřadů závisí převážně na počtu zaměstnanců, dále na výši nájemného a nepatrně také na poloze úřadu vzhledem k oblasti. Aby tedy daňová správa dodržovala zásadu hospodárnosti, měla by zaměstnávat takový počet pracovníků, který bude co nejvíce odpovídat počtu daňových subjektů a dále uzavírat nájemní smlouvy s co nejnižším nájemným nebo vyřešit tuto situaci jiným způsobem (přemístění úřadu do budovy ve vlastnictví státu, kraje, obce).

Změna polohy úřadu, resp. přizpůsobení příslušné oblasti na optimum je spíše nereálná. Jednalo by se o časově, organizačně (nutná změna legislativy) a finančně nákladnou záležitost. Jistá snaha o územní reorganizaci ale existuje. Za účelem synchronizace územní působnosti finančních úřadů s obcemi s rozšířenou působností byly vydány novely zákona č.531/1992 Sb., o územních finančních orgánech, platné od 1.1.2006 a 1.1.2007, které jisté přizpůsobení oblastí uzákoňují a zároveň snižují počet finančních úřadů.

Seznam použité literatury:

[1] K.Zvára: Regresní analýza. Academia, Praha 1989.

[2] K.Zvára, J.Štěpán: Pravděpodobnost a matematická statistika. MATFYZPRESS, Praha 1997.

[3] J.Anděl: Základy matematické statistiky. MATFYZPRESS, Praha 2005

[4] R.Hindls, S.Hronová, J.Seger: Statistika pro ekonomy. Professional Publishing, Praha 2003.

[5] I.Chvátalová, H.Marková, T.Gřivna: Základy veřejného práva. Oeconomica, Praha 2005.

[6] Zákon č.531/1992 Sb., o územních finančních orgánech, ve znění pozdějších předpisů.

[7] Zákon č.337/1992 Sb., o správě daní a poplatků, ve znění pozdějších předpisů.

[8] http://www.czso.cz, oficiální stránky Českého statistického úřadu.

[9] http://www.mfcr.cz, oficiální stránky ministerstva financí

Přílohy

Rozmístění finančních úřadů ve Středočeském kraji a oblastí v jejich územní působnosti