UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Modelování vybraných vlastností jalovce obecného nízkého za použití kvantilové regrese

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Eva Fišerová, Ph.D.**
Vypracoval(a): **Ondřej Kašík**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Matematika–ekonomie se zaměřením na bankovnictví/pojišťovnictví
Forma studia: prezenční
Rok odevzdání: 2022

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Ondřej Kašík

Název práce: Modelování vybraných vlastností jalovce obecného nízkého za použití kvantilové regrese

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2022

Abstrakt: Bakalářská práce má za cíl shrnout teoretické základy kvantilové regrese a ukázat možnost jejího využití na praktickém příkladu. Teoretická část zahrnuje popis tvorby modelu kvantilové regrese, uvedení některých význačných vlastností regresních parametrů a hodnocení tohoto modelu pomocí prostředků inferenční statistiky. V praktické části jsou pak poznatky z předchozích kapitol systematicky využity k regresní analýze datového souboru obsahujícího údaje o jalovci obecném nízkem.

Klíčová slova: lineární regresní modely, kvantilová regrese, bootstrap, jalovec obecný nízký

Počet stran: 66

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Ondřej Kašík

Title: Modelling of Selected Attributes of Alpine Junipers Using Quantile Regression

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2022

Abstract: The bachelor thesis aims to summarize basic theoretical concepts of quantile regression and to illustrate its possible application with a practical example. The theoretical part of the thesis includes describing the process of creating a quantile regression model, listing some relevant properties of the quantile regression estimates, and quantile regression inference. The techniques developed in previous chapters are systematically summarized and used in an analysis of a set containing data about Juniperus communis in the practical part of the thesis.

Key words: linear regression models, quantile regression, bootstrap, Juniperus communis

Number of pages: 66

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne 15. 7. 2022 .....

podpis

Obsah

Poděkování

Rád bych poděkoval paní doc. RNDr. Evě Fišerové, Ph.D. za nezměrnou trpělivost a ochotu při vedení této práce, za čas, který mi věnovala během konzultací a za cenné připomínky zvyšující celkovou kvalitu práce. Dále bych chtěl poděkovat panu RNDr. Miroslavu Zeidlerovi, Ph.D. za poskytnutí dat pro analýzu v praktické části této práce.

Úvod

Kvantilová regrese se stejně jako jiné druhy regresní analýzy zabývá modelováním vztahů závislosti vysvětlované proměnné na vysvětlující, popřípadě skupině vysvětlujících, proměnných. V reálných aplikacích zpravidla není na základě hodnot vysvětlujících proměnných odhadována přímo vysvětlovaná proměnná, ale nějaká její vhodná charakteristika, která tuto proměnnou v určitém smyslu dobře vystihuje. V tradičním přístupu, jenž nabízí klasická lineární regrese, je touto charakteristikou podmíněná střední hodnota. Tato je obecně vnímána jako dobrý odhad celkového chování vysvětlované proměnné v závislosti na vysvětlujících proměnných. Tento způsob je intuitivní a nabízí možnost relativně snadné interpretovatelnosti regresních parametrů v modelu.

Na druhou stranu je lineární regrese znevýhodněna množstvím předpokladů, které jsou na její použití kladeny. Splnění těchto předpokladů zůstává povětšinou záležitostí teorie a v praktických situacích, ve kterých soubory dat obsahují četná odlehlá pozorování, rozdělení dat je různě zešikmené a je zjevně porušen předpoklad homoskedasticity, může být lineární regrese přímo zavádějící a nesplňuje tak svůj účel výstižného a holistického popisu regresní závislosti.

Kvantilová regrese nabízí poněkud jiný přístup. Zatímco průměr, respektive střední hodnota, je charakteristika snadno ovlivnitelná odlehlými pozorováními, může jako robustní alternativa pro popis centrální polohy dobře posloužit medián. Pokud je předmětem zájmu vývoj regresní závislosti v jiných polohách hodnot vysvětlované proměnné než okolo středu, je možno model přizpůsobit a namísto podmíněného mediánu odhadovat jiný podmíněný kvantil. Takto lze analyzovat vztahy regresní závislosti ve velkých či naopak malých hodnotách vysvětlované

proměnné, což je rys, který klasická lineární regrese nemá.

Cílem této práce je uvést základní přehled teorie týkající se kvantilové regrese, ujasnit podobnosti i rozdíly oproti klasické lineární regresi a demonstrovat možnost praktického využití.

Práce sestává ze dvou hlavních částí. Teoretickou část tvoří první čtyři kapitoly. V první kapitole jsou stručně shrnuty nejdůležitější poznatky pojící se s klasickou lineární regresí. V druhé kapitole je zaveden základní teoretický aparát potřebný pro konstrukci modelu kvantilové regrese. Třetí kapitola se blíže věnuje vybraným vlastnostem odhadů parametrů kvantilové regrese. Čtvrtá kapitola se zabývá možnostmi hodnocení modelu kvantilové regrese prostřednictvím statistické inference. Praktická část práce se nachází v páté kapitole. Jsou zde s využitím softwaru R analyzována data o jalovci obecném nízkém pocházející ze studie [11]. Smyslem praktické části je názorně předvést možnost využití poznatků nabytých v teoretické části pro regresní analýzu reálného datového souboru. Ukázkový kód v softwaru R je připojen v příloze k této práci.

Kapitola 1

Klasický lineární regresní model

Tato kapitola má za cíl podat ucelený souhrn základních poznatků o lineární regresi, což je nezbytný krok před přechodem ke kvantilové regresi jako takové. Obsah kapitoly zahrnuje tvorbu regresního modelu, způsob odhadu neznámých parametrů, testování hypotéz a konstrukci intervalových odhadů pro tyto parametry a hodnocení regresního modelu prostřednictvím koeficientu determinace. V poslední části této první kapitoly se zaměříme na vysvětlení přístupu k modelování v situaci, kdy máme regresní model, ve kterém se vyskytují kvalitativní proměnné.

Podkapitoly 1.1, 1.2 a 1.3 vycházejí z literatury [3], podkapitola 1.4 pak z [1].

1.1. Konstrukce lineárního modelu, jeho předpoklady a odhad neznámých parametrů

U klasické lineární regrese sledujeme závislost podmíněné střední hodnoty závislé proměnné na hodnotě jedné či několika nezávislých proměnných. Model pro hodnotu závislé proměnné i-tého pozorování v modelu s p nezávislými proměnnými můžeme zapsat ve tvaru:

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \ i = 1, ..., n, \tag{1.1}$$

kde Y_i je vysvětlovaná (závislá) proměnná, $x_j, j = 1, ..., p$ jsou vysvětlující (nezávislé) proměnné, x_{ij} je hodnota j-té vysvětlující proměnné pro i-tý subjekt, β_j jsou odhadované parametry modelu, kterých je $p + 1$, n je počet pozorování

a ϵ_i jsou tzv. chybové členy (náhodné odchylky). Chybové členy jsou náhodné veličiny, o kterých se předpokládá:

$$E(\epsilon_i) = 0,\ var(\epsilon_i) = \sigma^2, i = 1, ..., n, \tag{1.2}$$

$$cov(\epsilon_i, \epsilon_j) = 0, \forall i \neq j; i, j = 1, ..., n, \tag{1.3}$$

tedy, že chyby mají nulovou střední hodnotu, stejný konečný nenulový rozptyl (tzv. předpoklad homoskedasticity) a jsou nekorelované.

Pro potřeby inferenční statistiky se obvykle přidává ještě předpoklad, že se chyby řídí normálním rozdělením, tedy, že

$$\epsilon_i \sim N(0, \sigma^2),\ i = 1, ..., n. \tag{1.4}$$

Model pro odhad podmíněné střední hodnoty závislé proměnné pak vypadá takto:

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip},\ i = 1, ..., n. \tag{1.5}$$

Neznámé parametry modelu odhadujeme metodou nejmenších čtverců:

$$arg_{\beta \in \mathbb{R}}\ min \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \tag{1.6}$$

Při odhadu neznámých parametrů lineárního regresního modelu tedy minimalizujeme součet druhých mocnin odchylek odhadnutých hodnot od naměřených pozorování.

1.2. Vektorový zápis modelu

Model (1.1) je možné zapsat také vektorově takto:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.7}$$

kde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_p \end{pmatrix},\ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}.$$

Předpoklady (1.2) pro lineární regresní model potom lze napsat vektorově takto:

$$E(\boldsymbol{\epsilon}) = \mathbf{0},\ var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}, \tag{1.8}$$

kde $\mathbf{0}$ označuje nulový vektor a $\boldsymbol{I}$ jednotkovou matici.

Pro odhad parametrů v tomto tvaru platí:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{1.9}$$

Tento odhad je nestranný, tedy:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{1.10}$$

a pro varianční matici tohoto odhadu platí:

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}. \tag{1.11}$$

Dále zaveďme veličinu S_e, což je tzv. reziduální součet čtverců definovaný:

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2, \tag{1.12}$$

kde

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip},\ i = 1, ..., n$$

jsou tzv. vyrovnané hodnoty a

$$e_i,\ i = 1, ..., n,$$

jsou tzv. rezidua.

Potom je statistika

$$S^2 = \frac{S_e}{n - (p+1)} \tag{1.13}$$

nestranným odhadem parametru σ^2.

1.3. Testování hypotéz o regresních parametrech, intervalové odhady těchto parametrů, koeficient determinace

K intervalovým odhadům a testování hypotéz o regresních parametrech slouží za předpokladu (1.4) testová statistika

$$T_j = \frac{\hat{\beta}_j - \beta_j}{S\sqrt{\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}_{j,j}}} \sim t_{n-(p+1)}, \quad (1.14)$$

kde $\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}_{j,j}$ značí $(j+1)$-ní diagonální prvek matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}, j = 0, ..., p$ a $t_{n-(p+1)}$ je Studentovo rozdělení s $n-(p+1)$ stupni volnosti.

Na základě statistiky (1.14) testujeme hypotézu $H_0 : \beta_j = 0$ oproti alternativě $H_A : \beta_j \neq 0$, tedy hypotézu, že j-tá nezávislá proměnná nemá na hodnotu závislé proměnné vliv oproti alternativě, že závislá proměnná závisí na j-té nezávislé proměnné

Na hladině významnosti $\alpha \in (0, 1)$ pak hypotézu H_0 zamítáme, pokud se statistika (1.14) realizuje hodnotou menší než je $-(1-\alpha/2)$-kvantil, anebo větší než $(1-\alpha/2)$-kvantil rozdělení $t_{n-(p+1)}$. Přesná definice kvantilu je uvedena v podkapitole 2.1.

Intervalový odhad pro jednotlivé regresní parametry vypadá takto:

$$\beta_j \in \langle \hat{\beta}_j - t_{n-(p+1),1-\alpha/2} S\sqrt{\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}_{j,j}}; \hat{\beta}_j + t_{n-(p+1),1-\alpha/2} S\sqrt{\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}_{j,j}} \rangle, \quad (1.15)$$

Pro posouzení kvality regresního modelu slouží koeficient determinace, definovaný

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} = 1 - \frac{S^2_{Y-\hat{Y}}}{S^2_Y}, \quad (1.16)$$

kde

$$S^2_Y = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad (1.17)$$

je tzv. celkový rozptyl,

$$S^2_{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \tag{1.18}$$

je rozptyl vyrovnaných veličin a

$$S^2_{Y-\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}S_e \tag{1.19}$$

je rozptyl reziduí.

Mezi těmito veličinami platí vztah

$$S^2_Y = S^2_{\hat{Y}} + S^2_{Y-\hat{Y}}. \tag{1.20}$$

Hodnota koeficientu determinace $R \in <0;1>$ vyjadřuje, jakou část rozptylu vysvětlované proměnné se podařilo vysvětlit pomocí regresního modelu. Čím větší hodnota koeficientu determinace, tím větší část rozptylu je vysvětlena.

Hodnotu koeficientu determinace však také ovlivňuje počet parametrů regresní funkce. S rostoucím počtem regresních parametrů roste také koeficient determinace. Z toho důvodu se zavádí upravený koeficient determinace:

$$R^2_a = 1 - \frac{(n-1)S^2_{Y-\hat{Y}}}{(n-p-1)S^2_Y} \tag{1.21}$$

Pro velká n jsou však rozdíly mezi hodnotami R^2 a R^2_a prakticky zanedbatelné.

Koeficient determinace pak lze použít k testování hypotézy

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \tag{1.22}$$

oproti alternativě

$$H_A : \exists j,\ j = 1, ..., p : \beta_j \neq 0. \tag{1.23}$$

Poté opět za předpokladu (1.4) má testová statistika

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p} \tag{1.24}$$

Fisherovo-Snedecorovo rozdělení pravděpodobnosti s počtem regresních parametrů p a počtem stupňů volnosti $n-p-1$, označujeme $F_{p,n-p-1}$. Hypotézu H_0, že všechny regresní koeficienty jsou nulové, zamítáme na hladině statistické významnosti α, pokud hodnota testové statistiky (1.24) překročí $(1-\alpha)$-kvantil rozdělení $F_{p,n-p-1}$.

1.4. Regresní model s kategoriálními proměnnými

Vysvětlující proměnné v regresním modelu mohou být nejen kvantitativního, ale i kvalitativního (kategoriálního) charakteru. Takovými proměnnými mohou být například pohlaví, rasa, státní příslušnost či vzdělání. V takové situaci se k modelování užívá tzv. umělých proměnných. Pro jednoduchost si techniku umělých proměnných uveďme na modelu s pouze jednou kvantitativní a jednou kategoriální proměnnou. V případě modelu s více proměnnými by se postupovalo analogicky

Mějme model se závislou proměnnou Y, kterou modelujeme pomocí kvantitativní vysvětlující proměnné x a kategoriální proměnné z, která nabývá m hodnot (kategorií). Jedna z těchto m kategorií je považována za výchozí (referenční), nechť je jí kupříkladu kategorie *1*. Základní model pak má tvar:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{[z_2]}(i) + ... + \beta_m I_{[z_m]}(i) + \epsilon_i,\ i = 1, ..., n, \tag{1.25}$$

kde I je indikátorová funkce definovaná jako

$$I_{[z_k]}(i) = \begin{cases} 1 & i\text{-té pozorování má hodnotu kategoriální proměnné } z \text{ rovnu } k, \\ & k = 2, ..., m \\ 0 & jinak \end{cases} \tag{1.26}$$

Může se také stát, že je vliv kvantitativní vysvětlující proměnné x na vysvětlovanou proměnnou Y odlišný u různých kategorií kvalitativní vysvětlující proměnné z. Proměnné x a z jsou pak v takzvané interakci a regresní model můžeme zapsat:

$$\begin{aligned} Y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{[z_2]}(i) + ... + \beta_m I_{[z_m]}(i) + \beta_{m+1} x_i I_{[z_2]}(i) + ...+ \\ + \beta_{2m-1} x_i I_{[z_m]}(i) + \epsilon_i,\ i = 1, ..., n. \end{aligned} \tag{1.27}$$

V regresním modelu se tak oproti předchozímu případu navíc objeví členy s interakcemi.

Všechna tvrzení pro lineární regresi uvedená v této kapitole zůstávají platná. Kategoriální proměnné vlastně pouze rozloží regresní model na m podmodelů (pro každou kategorii jeden). Každý podmodel pak má své vlastní hodnoty regresních parametrů:

$$
\begin{aligned}
Y_i =& \beta_0 + \beta_1 x_i + \epsilon_i && pro\ z_1 \\
Y_i =& (\beta_0 + \beta_2) + (\beta_1 + \beta_{m+1}) \cdot x_i + \epsilon_i && pro\ z_2 \\
& \vdots \\
Y_i =& (\beta_0 + \beta_m) + (\beta_1 + \beta_{2m-1}) \cdot x_i + \epsilon_i && pro\ z_m
\end{aligned}
$$

Kapitola 2

Úvod do kvantilové regrese

V této kapitole se již zaměříme na základní myšlenky a postupy související s kvantilovou regresí. Nejprve zadefinujeme nejdůležitější pojmy s kvantilovou regresí spojené, kterými jsou kvantil, kvantilová funkce a podmíněna kvantilová funkce. Dále uvedeme podobu modelu kvantilové regrese, způsob odhadu jeho parametrů a v čem se teto odhad liší oproti situaci v klasické lineární regresi. Na závěr si vysvětlíme, jak interpretovat regresní parametry v nejjednodušším případě modelu kvantilové regrese s jednou vysvětlující proměnnou kvantitativního charakteru.

K sepsání kapitoly bylo užito zdrojů [3], [8], [9] a [10].

2.1. Kvantily a kvantilová funkce

Náhodná veličina X je určena svou distribuční funkcí F_X definovanou jako

$$F_X(x) = P(X \leq x),\ x \in \mathbb{R}, \tag{2.1}$$

kde $P(X \leq x)$ značí pravděpodobnost, že se náhodná veličina X realizuje hodnotou menší než nebo rovnou x.

τ-kvantil náhodné veličiny X je pak definován následovně:

Definice 1 *Nechť $\tau \in (0,1)$. τ-kvantil náhodné veličiny X je takové reálné číslo x_τ pro které platí*

$$P(X \leq x_\tau) \geq \tau \tag{2.2}$$

a současně

$$P(X \geq x_\tau) \geq 1 - \tau. \tag{2.3}$$

Uvažujme nyní spojitou náhodnou veličinu X s distribuční funkcí F_X, která je spojitá a rostoucí všude tam, kde $0 < F_X < 1$. Pak je τ-kvantil x_τ náhodné veličiny X určen jednoznačně jako

$$F_X(x_\tau) = \tau \tag{2.4}$$

a tedy

$$x_\tau = {F_X}^{-1}(\tau) = Q_X(\tau). \tag{2.5}$$

Funkce $Q(\tau)$ se nazývá kvantilová funkce a je inverzní funkcí k distribuční funkci příslušné náhodné veličiny.

Toto se týká spojitých náhodných veličin se spojitou a rostoucí distribuční funkcí. Příkladem je kvantilová funkce normovaného normálního rozdělení (obrázek 2.1) Obecně však není τ-kvantil určen jednoznačně. Pro diskrétní náhodnou veličinu může existovat několik nebo dokonce celý ohraničený interval hodnot, které splňují podmínky uvedené v definici. Abychom zajistili jednoznačnost kvantilu, je možné kvantilovou funkci předefinovat následovně:

Definice 2 *Nechť X je náhodná veličina s distribuční funkcí $F_X(x)$ a nechť $\tau \in (0,1)$. Pak funkci*

$$Q_X(\tau) = {F_X}^{-1}(\tau) = inf\{x \in \mathbb{R} : F_X(x) \geq \tau\} \tag{2.6}$$

nazýváme kvantilovou funkcí a číslo $x_\tau = Q_X(\tau)$ nazýváme τ-kvantilem rozdělení s distribuční funkcí $F_X(x)$.

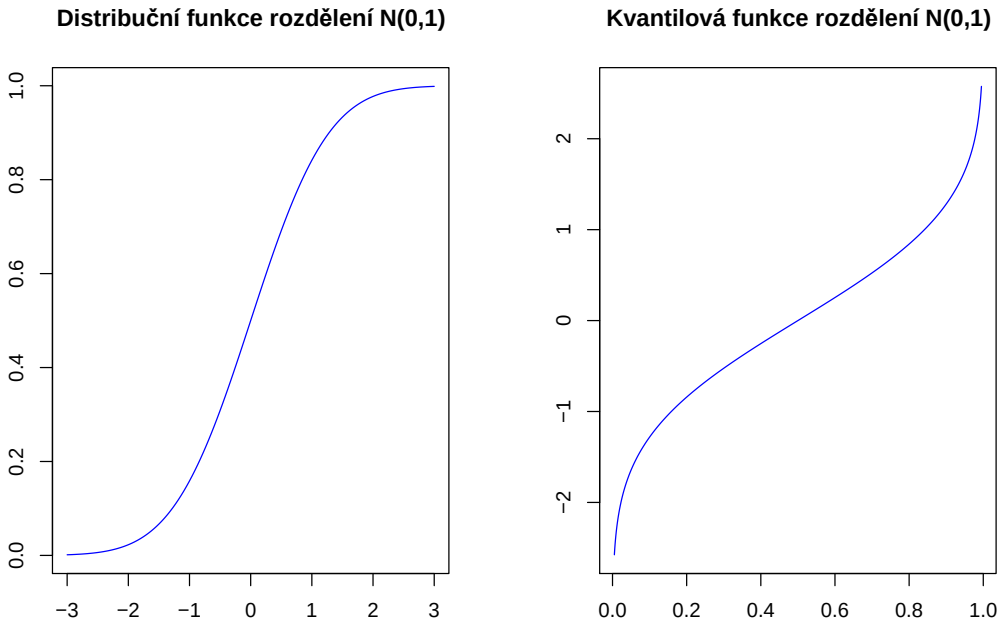
Některé kvantily mají speciální názvy:

$x_{0,5}$ se nazývá medián; $x_{0,25}$ dolní kvartil; $x_{0,75}$ horní kvartil;

$x_{\frac{k}{10}}, k = 1, 2, ..., 9$, je tzv. k-tý decil;

$x_{\frac{k}{100}}, k = 1, 2, ..., 99$, je tzv. k-tý percentil.

Než přejdeme k modelu kvantilové regrese, je třeba ještě definovat podmíněnou kvantilovou funkci. Nejprve budeme definovat podmíněnou distribuční funkci, poté podmíněnou kvantilovou funkci.



Obrázek 2.1: Distribuční a kvantilová funkce normovaného normálního rozdělení

Definice 3 *Nechť pro každou borelovskou množinu S a pro každé $y \in \mathbb{R}$ existuje funkce $F_{Y|X}(y|x)$ taková, že platí*

$$P(Y \leq y|X \in S) = \int_S F_{Y|X}(y|x)dF_X(x). \tag{2.7}$$

Potom funkci $F_{Y|X}(y|x)$ nazveme podmíněnou distribuční funkcí náhodné veličiny Y při daném $X \in S$.

Definice 4 *Nechť $(X,Y)'$ je náhodný vektor definovaný na pravděpodobnostním prostoru a $F_{Y|X}(y|x)$ je podmíněná distribuční funkce náhodné veličiny Y při daném $X = x$. Pak podmíněná kvantilová funkce náhodné veličiny Y při daném $X = x$ je definována vztahem*

$$Q_{Y|X}(\tau|x) = F^{-1}_{Y|X}(\tau|x) = inf\{x \in \mathbb{R} : F_{Y|X}(y|x) \geq \tau\},\ x \in \mathbb{R},\ \tau \in (0,1). \tag{2.8}$$

2.2. Model kvantilové regrese a odhad jeho parametrů

Kvantilová regrese se liší od klasické lineární regrese tím, že místo podmíněné střední hodnoty odhadujeme podmíněný medián (tzv. mediánová regrese), popřípadě libovolný jiný podmíněný kvantil vysvětlované proměnné.

Podobně jako v klasickém lineárním modelu uvažujme, že proměnná Y závisí na hodnotách vysvětlujících proměnných:

$$Y_i = \beta_0(\tau) + \beta_1(\tau)x_{i1} + ... + \beta_p(\tau)x_{ip} + \epsilon_i(\tau),\ i = 1, ..., n. \tag{2.9}$$

Parametry modelu β jsou nyní funkcí příslušného kvantilu τ. Předpokládáme, že chybové členy ϵ mají τ-kvantil roven nule. Asymptotický postup pro intervalové odhady a testování hypotéz o parametrech kvantilové regrese pak vychází z předpokladu, že jsou chybové členy nezávislé a stejně rozdělené.

Model pro odhad podmíněného τ-kvantilu má pak tuto podobu:

$$Q_{Y|X}(\tau|x_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + ... + \beta_p(\tau)x_{ip},\ i = 1, ..., n. \tag{2.10}$$

Odhad parametrů tohoto modelu se provádí řešením následujícího minimalizačního problému

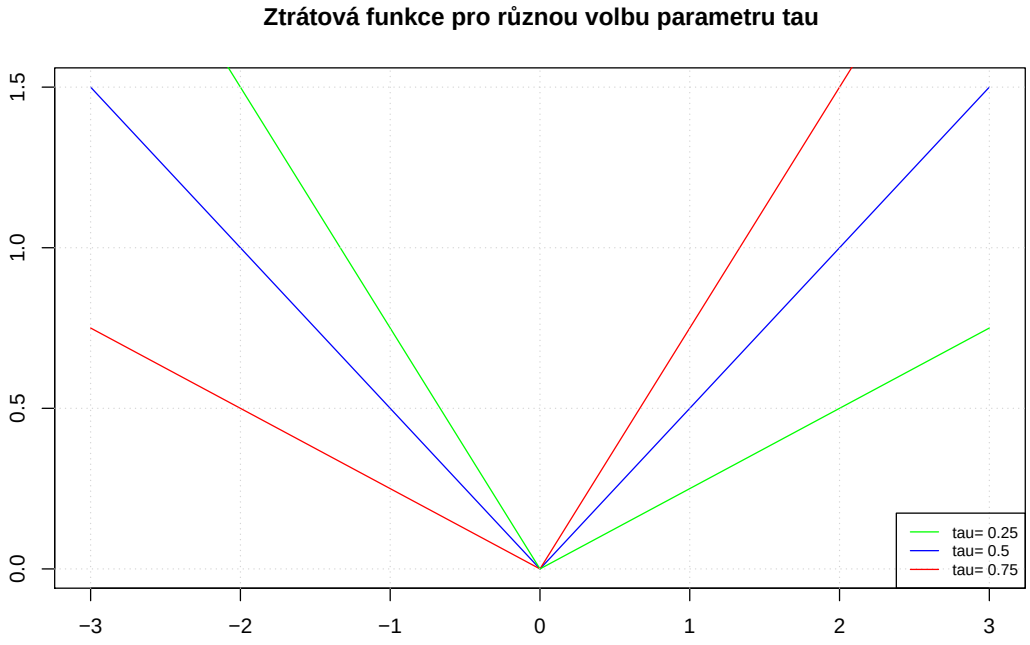
$$arg_{\beta\in\mathbb{R}}\ min \sum_{i=1}^{n} \rho_\tau\Big(Y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau)\Big). \tag{2.11}$$

Funkce $\rho_\tau(x)$ se nazývá ztrátová funkce a je definována takto

$$\rho_\tau(x) = \begin{cases} \tau|x| & x \geq 0 \\ (1-\tau)|x| & x < 0. \end{cases} \tag{2.12}$$

Graf ztrátové funkce pro vybrané kvantily se pak nachází na obrázku 2.2.

Neznámé parametry jsou tedy odhadovány minimalizací součtu vážených absolutních odchylek odhadovaných hodnot od naměřených pozorování. Pro každý τ-kvantil jsou řešením minimalizačního problému (2.11) jiné hodnoty odhadnutých parametrů.



Obrázek 2.2: Ztrátová funkce

Na rozdíl od lineární regrese nelze řešit problém (2.11) analyticky, jelikož ztrátová funkce není na celém svém definičním oboru diferencovatelná (konkrétně v bodech, kde je jedno či více reziduí rovno nule). Úlohu (2.11) je ovšem možné řešit jako minimalizační problém pomocí technik lineárního programování.

Pro model kvantilové regrese ve vektorovém tvaru

$$Q_{Y|X}(\tau|\boldsymbol{x}) = \mathbf{X}\boldsymbol{\beta}(\tau) + \boldsymbol{\epsilon}(\tau), \tag{2.13}$$

lze minimalizační úlohu (2.11) přepsat do podoby

$$\begin{array}{l}\min\{\tau\mathbf{1}_n^T\boldsymbol{\epsilon}_+ + (1-\tau)\mathbf{1}_n^T(-\boldsymbol{\epsilon})_+ | \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_+ - (-\boldsymbol{\epsilon})_+, \tau \in (0,1), \\ \boldsymbol{\beta} \in \mathbb{R}^{p+1}, [\boldsymbol{\epsilon}_+^T, (-\boldsymbol{\epsilon})_+^T]^T \in \mathbb{R}_+^n\}\end{array}, \tag{2.14}$$

kde

$$\boldsymbol{\epsilon}_+ = \begin{cases} \boldsymbol{\epsilon} & \boldsymbol{\epsilon} \geq 0 \\ 0 & \text{jinak} \end{cases}$$

a

$$(-\boldsymbol{\epsilon})_+ = \begin{cases} \boldsymbol{\epsilon} & -\boldsymbol{\epsilon} < 0 \\ 0 & \text{jinak} \end{cases}$$

Tuto úlohu lze vyjádřit v rovnicovém tvaru

$$arg_{\boldsymbol{x}} min\{\mathbf{c^T x}|\mathbf{Ax} = \mathbf{b}\} \tag{2.15}$$

kde

$$\begin{aligned}
\boldsymbol{c} &= (\mathbf{0}_{p+1}^T, \tau\mathbf{1}_n^T, (1-\tau)\mathbf{1}_n^T)^T, \\
\boldsymbol{x} &= (\boldsymbol{\beta}^T, \boldsymbol{\epsilon}_+^T, (-\boldsymbol{\epsilon})_+^T)^T, \\
\boldsymbol{A} &= [\boldsymbol{X}, \boldsymbol{I}, -\boldsymbol{I}], \\
\boldsymbol{b} &= \boldsymbol{Y}^T
\end{aligned}$$

2.3. Interpretace parametrů kvantilové regrese

Uvažujeme-li model kvantilové regrese (2.10), potom můžeme parametr $\beta_j(\tau)$ interpretovat jako hodnotu, o kterou se změní τ-kvantil podmíněného pravděpodobnostního rozdělení závislé proměnné, změní-li se j-tá nezávislá proměnná o jednotku. Tzn.:

$$\beta_j(\tau) = \frac{\partial Q_{Y|X}(\tau|x)}{\partial x_j} \tag{2.16}$$

Tato interpretace platí však pouze v případě, kdy při změně j-té vysvětlující proměnné o jednotku zůstanou hodnoty ostatních vysvětlujících proměnných zachovány. Neuvažujeme tedy multikolinearitu, tedy lineární závislost mezi jednotlivými vysvětlujícími proměnnými. Můžeme ovšem říct, že tato interpretace je jednoznačně platná pro lineární model kvantilové regrese s jednou vysvětlující proměnnou.

Kapitola 3

Vlastnosti odhadů parametrů kvantilové regrese

Cílem této kapitoly je uvést vybrané vlastnosti odhadů parametrů kvantilové regrese a představit si některé z těchto vlastností plynoucí výhody kvantilové regrese ve srovnání s klasickou lineární regresí.

V této kapitole se nebudeme zabývat asymptotickými vlastnostmi kvantilové regrese, jejichž analýza by přesahovala rámec této práce. Zájemcům o tuto problematiku lze doporučit kupříkladu literaturu [8].

Text kapitoly čerpá ze zdrojů [2] a [8].

3.1. Invariance vůči základním transformacím modelu

Jednou z užitečných vlastností kvantilové regrese je invariance vůči lineárním transformacím modelu. Tuto vlastnost má společnou s klasickým lineárním regresním modelem. U klasické lineární regrese tato vlastnost vyplývá z vlastností střední hodnoty. Přičteme-li k vysvětlované proměnné Y konstantu, změní se o tuto konstantu i odhad podmíněné střední hodnoty. Stejně tak, pokud proměnnou Y vynásobíme konstantou, je i odhad podmíněné střední hodnoty této proměnné roven součinu podmíněné střední hodnoty původní proměnné Y a této konstanty.

Symbolicky zapsáno:

$$E(a + cY|x) = a + cE(Y|x),\ a \in \mathbb{R},\ c \in \mathbb{R}. \tag{3.1}$$

Uvažujme model kvantilové regrese ve tvaru

$$Q_{Y|X}(\tau|\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{\beta}(\tau). \tag{3.2}$$

Odhad vektoru regresních parametrů $\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X})$ je funkcí příslušného kvantilu τ, vektoru hodnot vysvětlované proměnné $\boldsymbol{Y}$ a matice hodnot vysvětlujících proměnných $\boldsymbol{X}$. Nechť je dále $\boldsymbol{A}_{(p+1)\times(p+1)}$ regulární matice, $a > 0$ skalár a $\boldsymbol{c} \in \mathbb{R}^{p+1}$ vektor. Odhad $\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X})$ má potom tyto vlastnosti:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\tau, a\boldsymbol{Y}, \boldsymbol{X}) &= a\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X}),\\
\hat{\boldsymbol{\beta}}(\tau, -a\boldsymbol{Y}, \boldsymbol{X}) &= a\hat{\boldsymbol{\beta}}(1-\tau, \boldsymbol{Y}, \boldsymbol{X}),\\
\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y} + \boldsymbol{c}\boldsymbol{X}, \boldsymbol{X}) &= \hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X}) + \boldsymbol{c},\\
\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X}\boldsymbol{A}) &= \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X}).
\end{aligned}$$

Tyto vlastnosti plynou z vlastností podmíněné kvantilové funkce, viz. [2].

3.2. Zachování transformace vysvětlované proměnné

Někdy nastává v regresní analýze situace, která vyžaduje použití nelineární monotónní transformace. Například v případě asymetrického rozdělení se často využívá logaritmická transformace dat.

Problémem lineární regrese je, že vlastnost zachování transformace vysvětlované proměnné nemá. Transformujeme-li tedy závisle proměnnou Y monotónní funkcí h a odhadneme parametry tohoto transformovaného modelu, nezískáme parametry původního modelu jednoduše aplikací inverzní transformace na odhady parametrů v transformovaném modelu. Vyjádřeno symbolicky:

$$E(h(Y)|X) \neq h(E(Y|X)), \tag{3.3}$$

a tedy

$$E(Y|X) \neq h^{-1}(E(h(Y)|X)). \tag{3.4}$$

Velkou výhodou kvantilové regrese je, že na rozdíl od lineárního regresního modelu tuto výhodnou vlastnost má. V modelu kvantilové regrese totiž neodhadujeme podmíněnou střední hodnotu, ale podmíněnou kvantilovou funkci a pro tu platí

$$Q_{h(Y)|X}(\tau, \boldsymbol{x}) = h(Q_{Y|X}(\tau, \boldsymbol{x})). \tag{3.5}$$

Parametry původního modelu tak můžeme získat použitím inverzní transformace na parametry transformovaného modelu. Pro model (3.2) tedy můžeme psát

$$\hat{\boldsymbol{\beta}}(\tau, h(\boldsymbol{Y}), \boldsymbol{X}) = h(\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X})) \tag{3.6}$$

a

$$\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X}) = h^{-1}(\hat{\boldsymbol{\beta}}(\tau, h(\boldsymbol{Y}), \boldsymbol{X})). \tag{3.7}$$

Například uvažujme model, kde jako transformující funkci použijeme přirozený logaritmus. Potom

$$\hat{\boldsymbol{\beta}}(\tau, ln(\boldsymbol{Y}), \boldsymbol{X}) = ln(\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X})) \tag{3.8}$$

a pro parametry původního modelu platí

$$\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{Y}, \boldsymbol{X}) = e^{\hat{\boldsymbol{\beta}}(\tau, ln(\boldsymbol{Y}), \boldsymbol{X})}. \tag{3.9}$$

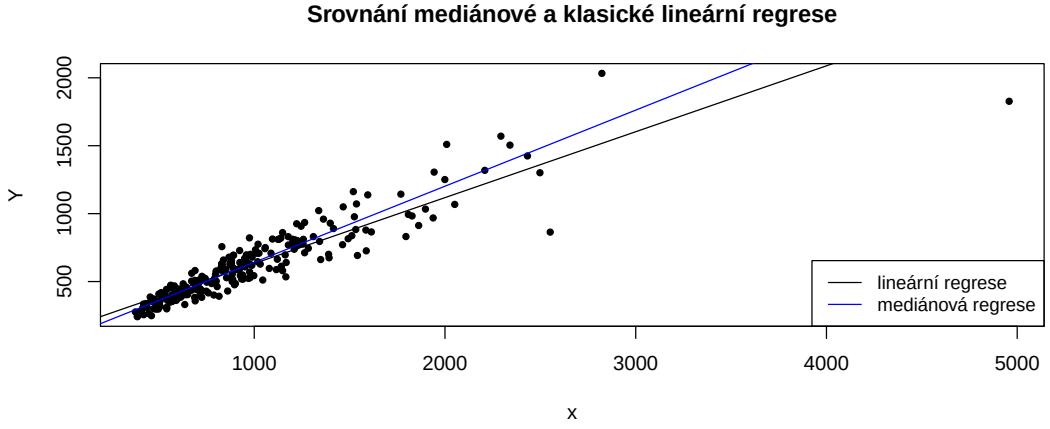
3.3. Robustnost

Jak již bylo zmíněno v 1. kapitole, existují určité předpoklady na rozdělení vysvětlované proměnné a jeho tvar. Porušení těchto předpokladů může výrazně snižovat kvalitu modelu, a tudíž i jeho vypovídací schopnost.

Robustnost je možno chápat jako jistou necitlivost vůči odlehlým pozorováním. Odlehlé pozorování (anglicky *outlier*) je v tomto kontextu chápáno jako taková hodnota vysvětlované proměnné, která se výrazně odchyluje od většiny ostatních hodnot. V lineárním regresním modelu může jediné takové odlehlé pozorování způsobit významné vychýlení regresní křivky. To vyplývá opět z vlastnosti střední

hodnoty, což je charakteristika citlivá na odlehlá pozorování a není tedy robustní. Tento jev může ovšem výrazně zkreslit interpretaci regresního modelu a je to jedna z nevýhod lineární regrese.

V případě kvantilové regrese nás ovšem nezajímá odhad podmíněné střední hodnoty nýbrž odhad podmíněných kvantilů. Jelikož velkou předností kvantilové regrese je právě modelování celého podmíněného kvantilového rozdělení vysvětlované veličiny, zajímá nás při modelování často právě vývoj v extrémech. Eliminace odlehlých pozorování z datového souboru při jeho analýze tedy není vhodným řešením, neboť právě situace v okolí těchto odlehlých pozorování je předmětem našeho zájmu. Naštěstí je právě kvantil robustní charakteristika, a proto je model kvantilové regrese vhodný i pro práci s daty obsahujícími odlehlá pozorování. Z vlastnosti ztrátové funkce (2.12) totiž vyplývá, že změna hodnoty vysvětlované veličiny u daného pozorování nezpůsobí změnu regresní křivky, dokud je zachováno znaménko rezidua, jinými slovy, dokud hodnota pozorování zůstává ve stejné polorovině od regresní křivky (nad křivkou anebo naopak pod ní).



Obrázek 3.1: Srovnání vlivu odlehlých pozorování na mediánovou a klasickou lineární regresi

Graf na obrázku 3.1 je ukázkou robustnosti mediánové regrese oproti klasické lineární regresi. Odlehlé pozorování v pravém horním rohu grafu vychyluje re-

gresní přímku lineární regrese, zatímco přímka pro mediánovou regresi zachovává směrnici určenou převážnou většinou dat. Graf byl vykreslen v softwaru R na základě dat z datového souboru *Engel*, který je součástí balíčku [7].

Kromě této vlastnosti, je kvantilová regrese velmi užitečná i v jiném ohledu. V lineárním regresním modelu stojí veškeré intervalové odhady a testování hypotéz na předpokladu normality. Kvantilová regrese je, na druhou stranu, indiferentní vůči tvaru rozdělení vysvětlované proměnné. Intervalové odhady a testování hypotéz v modelu kvantilové regrese je tedy možno provádět bez jakýchkoli omezujících předpokladů na tvar rozdělení.

Kapitola 4

Hodnocení modelu kvantilové regrese

V této kapitole se zaměříme na možné přístupy ke konstrukci intervalových odhadů regresních parametrů v modelech kvantilové regrese. Také si popíšeme základní princip bootstrapových metod a jejich výhodu v porovnání s asymptotickým přístupem. V závěru kapitoly představíme obdobu koeficientu determinace pro kvantilovou regresi.

Jako podklad pro sepsání této kapitoly posloužila literatura [2] a [8]. V těchto pracích lze dohledat také podrobnosti týkající se asymptotického odhadu varianční matice odhadů regresních parametrů.

4.1. Asymptotický odhad intervalů spolehlivosti regresních parametrů

Ve 2. kapitole je uveden postup výpočtu bodových odhadů regresních parametrů β. Tyto odhady jsou pro konstrukci regresního modelu nezbytné, avšak pro jeho posouzení nemají dostatečnou vypovídací hodnotu. Nesdělují nám totiž míru nejistoty, která je s těmito odhady spojena. Proto jsou zpravidla předmětem zájmu také intervalové odhady regresních parametrů.

V podkapitole 1.3 je popsána konstrukce intervalových odhadů parametrů lineární regrese. V kvantilové regresi, stejně jako v klasické lineární regresi, potřebujeme pro konstrukci intervalových odhadů kromě samotných bodových od-

hadů regresních parametrů také odhady jejich směrodatných odchylek. Ty získáme jako odmocniny příslušných diagonálních prvků asymptotické varianční matice odhadů regresních parametrů.

Problém nastává při samotném odhadu varianční matice. Podobně jako v klasické lineární regresi, i v případě kvantilové regrese dosti záleží na tom, zda jsou chybové členy stejně a nezávisle rozděleny či nikoliv.

V případě, že můžeme o chybových členech předpokládat stejné a nezávislé rozdělení, lze odhadovat varianční matici asymptoticky. Možný postup je uveden např. v [2].

Oboustranný interval spolehlivosti regresního parametru β_j pro τ kvantil na hladině významnosti α má podobu

$$\beta_j(\tau) \in (\hat{\beta}_j(\tau) - \hat{\sigma}_{\hat{\beta}_j(\tau)} z_{1-\alpha/2}; \hat{\beta}_j(\tau) + \hat{\sigma}_{\hat{\beta}_j(\tau)} z_{1-\alpha/2}), \tag{4.1}$$

kde $\hat{\sigma}_{\hat{\beta}_j(\tau)}$ je odhad směrodatné odchylky odhadu parametru $\beta_j(\tau)$ a $z_{1-\alpha/2}$ je $1-\alpha/2$ kvantil normovaného normálního rozdělení.

Pro malé n je výhodnější použít aproximaci Studentovým rozdělením

$$\beta_j(\tau) \in (\hat{\beta}_j(\tau) - \hat{\sigma}_{\hat{\beta}_j(\tau)} t_{n-(p+1),1-\alpha/2}; \hat{\beta}_j(\tau) + \hat{\sigma}_{\hat{\beta}_j(\tau)} t_{n-(p+1),1-\alpha/2}), \tag{4.2}$$

V praxi je však předpoklad nezávisle a stejně rozdělené chybové složky mnohdy porušen, což činí z použití asymptotiky nevhodný způsob odhadu. Jako alternativní přístup lze užít bootstrap, který neklade žádné nároky na rozdělení chybové složky.

4.2. Bootstrapová metoda pro kvantilovou regresi

Metoda bootstrap je zvláštním případem metod Monte Carlo a je užitečnou alternativou k asymptotickému přístupu pro odhadování kovarianční matice odhadů regresních parametrů. Obecně lze metodu bootstrap použít k odhadu různých charakteristik sledované náhodné veličiny jako jsou střední hodnota, medián nebo směrodatná odchylka.

V obecnosti funguje bootstrapová metoda tak, že z reálného souboru dat o rozsahu n, na jehož základě chceme odhadovat číselnou charakteristiku sledované náhodné veličiny, provedeme náhodný výběr s vracením o rozsahu n, tzn. každá z n hodnot z původního souboru se může ve výběru vyskytovat maximálně n-krát (ale nemusí být také do výběru vybrána vůbec). Odhad hledané číselné charakteristiky vypočítáme na základě dat v tomto výběru. Tento postup opakujeme M-krát. Celkem tedy budeme mít M náhodných výběrů, každý o rozsahu n, a M odhadů hledané číselné charakteristiky. Z těchto M odhadů utvoříme empirické rozdělení pravděpodobnosti, které se nazývá bootstrapové. Z tohoto rozdělení můžeme odhadovat střední hodnotu a směrodatnou odchylku hledané charakteristiky a na jejich základě zkonstruovat intervalový odhad.

V případě kvantilové regrese lze metodu bootstrap použít k odhadu kovarianční matice odhadů regresních parametrů nebo i přímo k odhadu intervalů spolehlivosti pro tyto parametry. Bootstrap metodu je možné použít také přímo k odhadu samotného rozdělení odhadů regresních parametrů.

Ilustrujme použití jedné z bootstrapových metod na jednoduchém příkladu kvantilové regrese s pouze jednou vysvětlující proměnnou x a sledovanou vysvětlovanou proměnnou y. Máme tak soubor obsahující n párů vysvětlující a vysvětlované proměnné $(x_i, y_i),\ i = 1, ..., n$. Poté se postupuje následovně:

---

1. Z $\{(x_i, y_i),\ i = 1, ..., n\}$ vybereme n párů s vracením. Tím získáme bootstrapový výběr párů $\{(x_j^*, y_j^*),\ j = 1, ..., n\}_b$.
2. Tento bootstrapový výběr dosadíme do vzorce (2.11) pro odhad parametrů modelu kvantilové regrese. Získáme tak bootstrapový odhad vektoru regresních parametrů $\hat{\boldsymbol{\beta}}_b^*(\tau)$.
3. Tento proces opakujeme M-krát a dostaneme M odhadů vektoru regresních parametrů.

Na základě těchto M bootstrapových odhadů jsme schopni vypočítat varianční matici odhadu vektoru regresních parametrů:

$$var(\hat{\boldsymbol{\beta}}^*(\tau)) = \frac{1}{M}\sum_{b=1}^{M}[\hat{\boldsymbol{\beta}}_b^*(\tau) - \overline{\hat{\boldsymbol{\beta}}}_b^*(\tau)][\hat{\boldsymbol{\beta}}_b^*(\tau) - \overline{\hat{\boldsymbol{\beta}}}_b^*(\tau)]^T, \tag{4.3}$$

kde $\overline{\hat{\boldsymbol{\beta}}}_b^*(\tau) = \frac{1}{M}\sum_{b=1}^{M}\hat{\beta}^*(\tau)$.

Intervalový odhad jednotlivých odhadů regresních parametrů pak můžeme konstruovat podle vzorců (4.1) nebo (4.2), kde za odhad směrodatné odchylky $\hat{\sigma}_{\hat{\beta}_j(\tau)}$ dosadíme odmocninu z j-tého diagonálního prvku matice $var(\hat{\boldsymbol{\beta}}^*(\tau))$.

Pro menší výběry je možné použít jednoduchý způsob, který se nazývá percentilový odhad intervalu spolehlivosti a vychází z kvantilu bootstrapového rozdělení odhadu daného parametru. Má podobu:

$$\beta_j(\tau) \in [\hat{\beta}^*_{j,\alpha/2}(\tau) \leq \hat{\beta}_j(\tau) \leq \hat{\beta}^*_{j,1-\alpha/2}(\tau)] \tag{4.4}$$

přičemž $\hat{\beta}^*_{j,\alpha/2}(\tau)$ se vypočítá jako $\alpha/2$ a $\hat{\beta}^*_{j,1-\alpha/2}(\tau)$ jako $1-\alpha/2$ výběrový kvantil bootstrapového odhadu j-tého regresního parametru.

4.3. Hodnocení kvality regresního modelu

Jak je uvedeno v kapitole 1, v případě klasického lineárního regresního modelu slouží k posouzení kvality tohoto modelu koeficient determinace, který udává tu část rozptylu vysvětlované proměnné y, již se podařilo vysvětlit regresním modelem.

Obdoba koeficientu determinace existuje i v případě kvantilové regrese. Rozdílem je, že se místo se součtem kvadratických odchylek pracuje se součtem vážených absolutních odchylek. Mějme tedy model kvantilové regrese ve tvaru

$$Q_{Y|X}(\tau|x_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + ... + \beta_p(\tau)x_{ip},\ i = 1, ..., n. \tag{4.5}$$

Potom je reziduální součet vážených absolutních odchylek roven

$$V^1(\tau) = \sum_{i=1}^{n}\rho_\tau(Y_i - \hat{\beta}_0(\tau) - \sum_{j=1}^{p}x_{ij}\hat{\beta}_j(\tau)), \tag{4.6}$$

tedy jde vlastně o účelovou funkci, jejíž minimalizací hledáme odhad neznámých parametrů regresního modelu.

Dále definujme celkový součet vážených absolutních odchylek jako

$$V^0(\tau) = \sum_{i=1}^{n}\rho_\tau(Y_i - \hat{\beta}_0(\tau)), \tag{4.7}$$

což je v podstatě reziduální součet vážených absolutních odchylek v nulovém modelu, v němž se nevyskytuje žádná vysvětlující proměnná, ale pouze konstanta. Pak lze definovat tzv. pseudokoeficient determinace (tedy analogii ke koeficientu determinace v prostředí kvantilové regrese) jako

$$R^2_{pseudo}(\tau) = 1 - \frac{V^1(\tau)}{V^0(\tau)}. \tag{4.8}$$

Protože $V^0(\tau) \geq V^1(\tau)$, platí $R^2_{pseudo}(\tau) \in <0;1>$.

Interpretace pseudokoeficientu determinace je obdobná jako u klasické lineární regrese, tedy rostoucí hodnota pseudokoeficientu determinace indikuje vhodnější model. Je však třeba zároveň brát na zřetel, že pseudokoeficient determinace není jedinou a určující charakteristikou kvality modelu. Přidávání dalších a dalších vysvětlujících proměnných pouze za účelem zvýšení hodnoty $R^2_{pseudo}(\tau)$ je tedy poměrně zavádějící přístup. Jelikož se hodnota $R^2_{pseudo}(\tau)$ liší v závislosti na volbě kvantilu τ, je vhodné pro posouzení celého modelu vzít v úvahu nějakou souhrnnou charakteristiku, kupříkladu průměr hodnot pseudokoeficientu determinace napříč několika různými kvantily.

Kapitola 5

Aplikace kvantilové regrese při analýze vlastností jalovce obecného nízkého

V této části si získané poznatky systematicky shrneme na příkladu statistické analýzy datového souboru, který obsahuje údaje o jalovci obecném nízkém. Sběr data byl uskutečněn ve vysokohorské nezalesněné části Hrubého Jeseníku v průběhu července roku 2017. Data byla poskytnuta z Katedry ekologie a životního prostředí Přírodovědecké fakulty Univerzity Palackého v Olomouci. Celý popis studie je k nalezení v [11].

Datový soubor čítá 326 jedinců. Vlastnosti, které v této práci budeme analyzovat jsou *výška*, *šířka* a *délka* jalovce, *nadmořská výška*, ve které se daný jalovec nachází, a jeho *pohlaví*. Vzhledem k tomu, že z jednotlivých rozměrů jalovce lze jednoznačně odlišit pouze výšku, vytvoříme novou proměnnou *tloušťka*, kterou získáme prostým součinem délky a šířky. Původní dvě proměnné *délka* a *šířka* tedy sloučíme do jedné jejich vynásobením. Nově proto dostáváme pouze proměnné *výška*, *tloušťka*, *nadmořská výška* a *pohlaví*.

Máme tedy tři kvantitativní a jednu kvalitativní proměnnou. Výška je uvedena v centimetrech, tloušťka v centimetrech čtverečních, nadmořská výška v metrech nad úrovní mořské hladiny. Proměnná *pohlaví* nabývá třech kategorií, kterými jsou *samčí*, *samičí* a *neurčité*. Příslušnost do dané kategorie je dána výskytem odpovídajících generativních orgánů (šištic). Pro zjednodušení budeme dále jed-

notlivé kategorie proměnné pohlaví označovat zkratkami M (*samčí*), F (*samičí*) a 0 (*neurčité*).

V regresní analýze zvolíme modelování tloušťky v závislosti na ostatních vlastnostech. Tloušťka se tak stává závisle proměnnou a ke zbylým vlastnostem budeme přistupovat jako k nezávislým. Cílem bude najít co nejvhodnější model pro tuto závislost. Faktory, které budeme brát při hodnocení kvality modelu v úvahu, zahrnují skutečnost, zda jsou jednotlivé regresní parametry statisticky významné, šířku jejich intervalů spolehlivosti a hodnotu pseudokoeficientu determinace. V ideálním modelu by měly být všechny parametry významné (tedy měly by být významné všechny v modelu obsažené vysvětlující proměnné), intervaly spolehlivosti jednotlivých regresních parametrů by měly být co nejužší a pseudokoeficient determinace zase co největší.

Je zřejmé, že v praxi málokdy narazíme na ideální model. Dílčí cíle, které jsme si výše stanovili, se navíc mohou vzájemně vylučovat. Nesmíme také zapomínat na fakt, že provádíme-li kvantilovou regresi, neodhadujeme podmíněnou střední hodnotu, zato můžeme odhadovat libovolný podmíněný kvantil. Pokud hodnotíme daný model kvantilové regrese, nezajímají nás jeho charakteristiky pouze při odhadu jednoho konkrétního podmíněného kvantilu, nýbrž napříč všemi kvantily. Zmíněné charakteristiky se samozřejmě při odhadu různých kvantilů liší, což činí situaci o něco komplikovanější. Proto se smíříme s tím, že v našem případě pravděpodobně nenajdeme model, který by byl dokonalý ve všech aspektech, a spokojíme se s určitým kompromisem.

Kromě hodnocení regresního modelu bude předmětem našeho zájmu také interpretace jeho výstupů. Budeme se zabývat odlišnostmi mezi jednotlivými kategoriemi pohlaví, srovnávat koeficienty klasické lineární a mediánové regrese, zkoumat vliv odlehlých pozorování a vše si pro snazší představu znázorníme graficky. Protože chceme využít plný potenciál kvantilové regrese, nebudeme se omezovat pouze na medián, ale podíváme se také, jaká je situace v extrémech, tzn. budeme modelovat závislost pro velmi malé i velmi velké kvantily.

Veškerá analýza bude provedena v softwaru R, který je volně dostupný ke sta-

žení na adrese <www.r-project.org>.

5.1. Popisná statistika

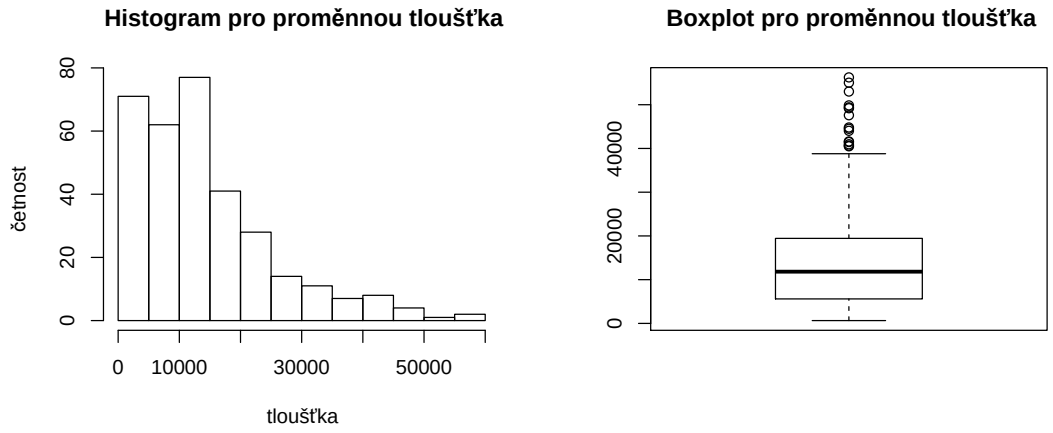
V tabulce 5.1 jsou uvedeny charakteristiky kvantitativních proměnných, které vystupují v modelu. Těmito proměnnými jsou tloušťka, výška a nadmořská výška a dále také proměnné vytvořené aplikací přirozeného logaritmu na tyto tři veličiny. U proměnné *tloušťka* si můžeme povšimnout dvou jevů, kterými jsou jednak velmi velká variabilita v této proměnné indikovaná velikostí směrodatné odchylky, jednak zešikmené rozdělení (značný rozdíl mezi mediánem a průměrem). Tento rys můžeme také vidět na obrázku 5.1. Na boxplotu rovněž vidíme další vlastnost vysvětlované proměnné, kterou je výskyt několika velkých odlehlých pozorování.

proměnná	dolní kvartil	medián	horní kvartil	průměr	směrodatná odchylka
tloušťka [cm^2]	5608	11828	19422,75	14192,64	11246,82
ln(tloušťka) [cm]	8,63	9,38	9,87	9,22	0,91
výška [cm]	153,25	216,00	303,00	237,67	121,27
ln(výška) [cm]	5,03	5,38	5,71	5,33	0,55
nadmořská výška [m.n.m.]	1330,77	1335,54	1357,47	1349,96	35,29
ln(nadmořská výška) [m.n.m.]	7,19	7,20	7,21	7,21	0,03

Tabulka 5.1: Popisná statistika kvantitativních proměnných užitých v analýze

Z těchto důvodů bude vhodné vysvětlovanou proměnnou transformovat pomocí přirozeného logaritmu a modelovat závislost pro tuto transformovanou veličinu. Logaritmus je monotónní funkce, můžeme tedy aplikovat poznatky z kapitoly 3.2. Srovnáme poté výsledky modelů pro původní a transformovanou proměnnou. Jak je vidět i z tabulky, použití logaritmu potlačuje velkou variabilitu proměnné *tloušťka*. Proto můžeme očekávat také při modelování lepší výsledky u transformované proměnné.

V tabulce 5.2 a na obrázku 5.2 je vidět rozložení některých číselných charakteristik proměnné *tloušťka* v rámci jednotlivých kategorií pohlaví. Zdá se, že samičí jalovce jsou nejtlustší, po nich následují samčí a nejtenčí bývají jedinci neurčitého pohlaví. Zda jsou tyto rozdíly signifikantní a projeví se i v modelu kvantilové regrese, zatím nemůžeme říct, nicméně máme alespoň důvod domnívat se, že rozdíly

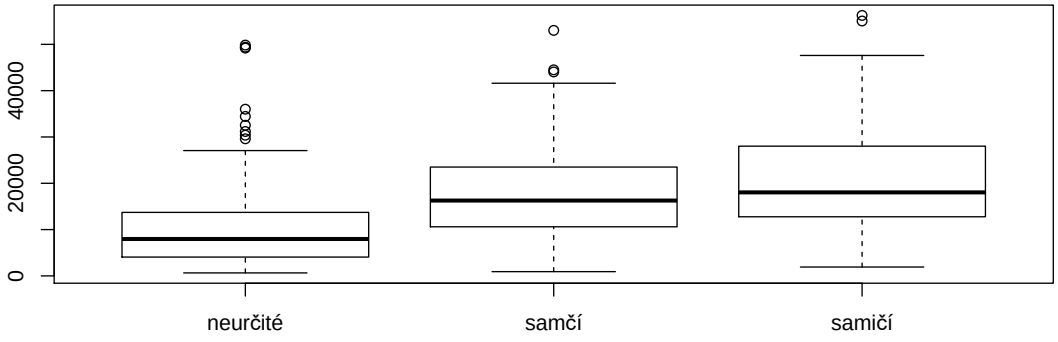


Obrázek 5.1: Histogram a boxplot pro proměnnou tloušťka

v tloušťce mezi pohlavími existují, a zahrnout do modelu *pohlaví* jako vysvětlující proměnnou.

pohlaví	dolní kvartil	medián	horní kvartil	průměr	směrodatná odchylka
neurčité	4076,25	7971,00	13706,25	10347,91	8805,15
samčí	10602,00	16268,00	23504,00	18670,38	12051,56
samičí	12768,00	18032,00	28012,00	21155,32	12011,83

Tabulka 5.2: Popisná statistika proměnné tloušťka v cm^2 v rámci jednotlivých kategorií pohlaví



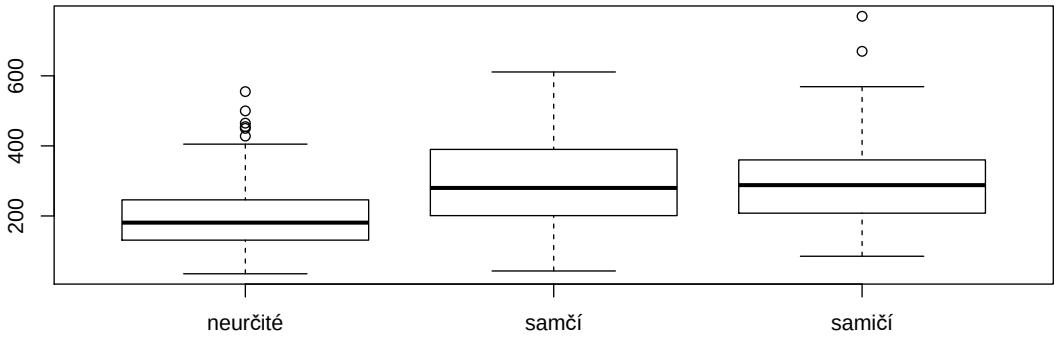
Obrázek 5.2: Boxplot pro proměnnou tloušťka v cm^2 v rámci jednotlivých kategorií pohlaví

V tabulce 5.3 a na obrázku 5.3 jsou uvedeny číselné charakteristiky pro

vysvětlující proměnnou výška. Seřazení podle kategorií zůstává vesměs stejné jako u tloušťky, přestože rozdíly mezi pohlavími jsou nyní výrazně menší. Můžeme si však povšimnout, že horní kvartil výšky je u samčích jalovců větší než u samičích.

pohlaví	dolní kvartil	medián	horní kvartil	průměr	směrodatná odchylka
neurčité	131,50	181,00	246,00	195,21	93,90
samčí	201,00	280,00	390,00	300,79	129,94
samičí	208,00	288,00	360,00	302,48	130,65

Tabulka 5.3: Popisná statistika proměnné výška v cm v rámci jednotlivých kategorií pohlaví

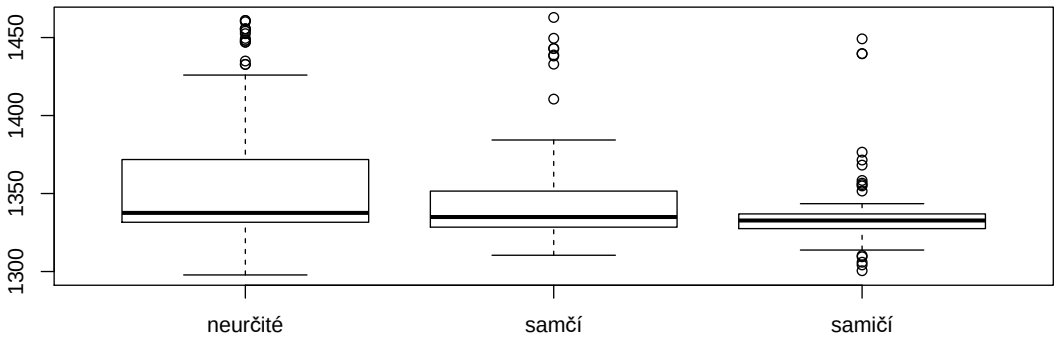


Obrázek 5.3: Boxplot pro proměnnou výška v cm v rámci jednotlivých kategorií pohlaví

Tabulka 5.4 a obrázek 5.4 udávají číselné charakteristiky nadmořské výšky pro jednotlivé kategorie pohlaví. Tato proměnná má oproti tloušťce a výšce nejmenší variabilitu a rozdíly mezi pohlavími jsou zde malé, ovšem seřazení jednotlivých kategorií je obrácené. Nejvýše jsou položeny bezpohlavní jalovce a ze všech kategorií pohlaví leží obecně v nejmenší nadmořské výšce samičí jalovce.

pohlaví	dolní kvartil	medián	horní kvartil	průměr	směrodatná odchylka
neurčité	1331,60	1337,59	1371,53	1354,23	35,95
samčí	1328,44	1334,94	1351,57	1350,41	38,74
samičí	1327,52	1332,72	1336,94	1337,44	26,68

Tabulka 5.4: Popisná statistika proměnné nadmořská výška v m.n.m v rámci jednotlivých kategorií pohlaví



Obrázek 5.4: Boxplot pro proměnnou nadmořská výška v m.n.m. rámci jednotlivých kategorií pohlaví

Na závěr se ještě podíváme na počty jedinců jednotlivých pohlaví. Jak vidíme v tabulce 5.5, nejvíce jedinců v souboru je tedy neurčitého pohlaví, zatímco menšinové samčí a samičí jalovce jsou zastoupeny v přibližně stejném množství.

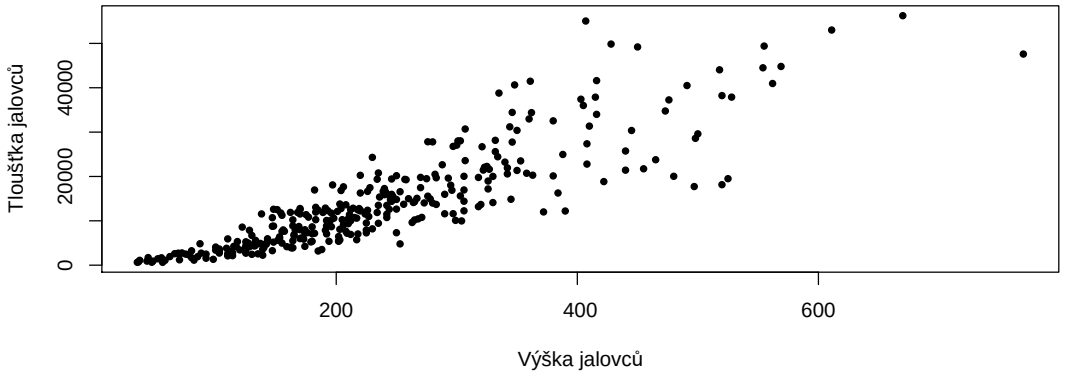
pohlaví	absolutní četnost	relativní četnost
neurčité	196	0,60
samčí	61	0,19
samičí	69	0,21

Tabulka 5.5: Zastoupení jalovců v souboru dle pohlaví

5.2. Model kvantilové regrese s jednou kvantitativní vysvětlující proměnnou

V této podkapitole si představíme základní funkce pro kvantilovou regresi v softwaru R. Vše si ukážeme na zjednodušeném modelu s pouze jednou vysvětlující proměnnou, kterou je kvantitativní proměnná *výška*. Vysvětlovanou proměnnou je kvantitativní proměnná *tloušťka*.

Na obrázku 5.5 vidíme bodový graf znázorňující hodnoty výšky a tloušťky jednotlivých jedinců v souboru. Jak by se dalo očekávat, zdá se, že s rostoucí hodnotou výšky se zvětšuje i tloušťka, mezi proměnnými by tedy měla existovat přímá závislost. Dále si lze všimnout zjevné heteroskedasticity v datech.



Obrázek 5.5: Bodový graf pro závislost tloušťky na výšce

Můžeme se pokusit data proložit přímkou, očekáváme tedy lineární závislost ve tvaru

$$tloušťka = \beta_0(\tau) + \beta_1(\tau) \cdot výška + \epsilon(\tau). \tag{5.1}$$

Abychom mohli používat funkce pro vytvoření a práci s modelem kvantilové regrese, je nejprve potřeba nahrát balíček Quantreg [7]

Základní funkcí pro vytvoření modelu kvantilové regrese je funkce

```
rq(formula,tau).
```

Funkce má dva základní parametry. Těmi jsou `formula`, což je předpis regresního modelu ve tvaru „závislá proměnná ∼ nezávislé proměnné“, tedy obdobně jako u příkazu `lm` pro klasickou lineární regresi. Oproti příkazu `lm` má však funkce `rq` navíc parametr `tau`. Tento parametr určuje kvantil, pro nějž mají být regresní parametry počítány. Defaultní hodnota tohoto parametru je `0.5`, tedy budou počítány parametry mediánové regrese.

Pro odhad parametrů modelu (5.1) použijeme předpis ve tvaru

```
model<-rq(tloustka~vyska,tau=...).
```

Například, chceme-li odhadovat parametry mediánové regrese, bude model vypadat takto:

```
model1<-rq(tloustka~vyska,tau=0.5).
```

Bodové odhady parametrů $\beta_0(0,5)$ a $\beta_1(0,5)$ získáme použitím příkazu

```
model1$coefficients
```

podobně jako v klasické regresi a dostaneme výstup:

```
(Intercept)       vyska
-5123.18861    79.61566 .
```

Bodové odhady regresních parametrů pro medián jsou tedy:

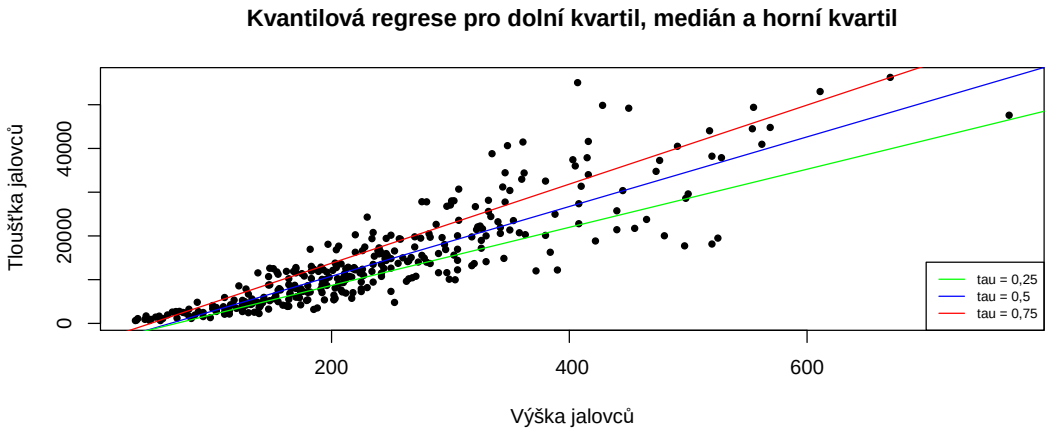
$$\hat{\beta}_0(0,5) = -5123,18861,\ \hat{\beta}_1(0,5) = 79,61566.$$

Analogicky bychom změnou parametru `tau` získali bodové odhady pro horní a dolní kvartil, popřípadě libovolný jiný kvantil. V tabulce 5.6 jsou uvedeny hodnoty regresních koeficientů pro dolní a horní kvartil, medián a klasickou lineární regresi. Hodnoty jsou zaokrouhleny na dvě desetinná místa. Jednotlivé regresní přímky pro medián, dolní a horní kvartil jsou vykresleny na obrázku 5.6. Na obrázku 5.7 jsou pak graficky srovnány regresní přímky pro lineární a mediánovou regresi. Je vidět, že v případě tohoto modelu a zadaných dat se lineární a mediánová regrese liší jen nepatrně.

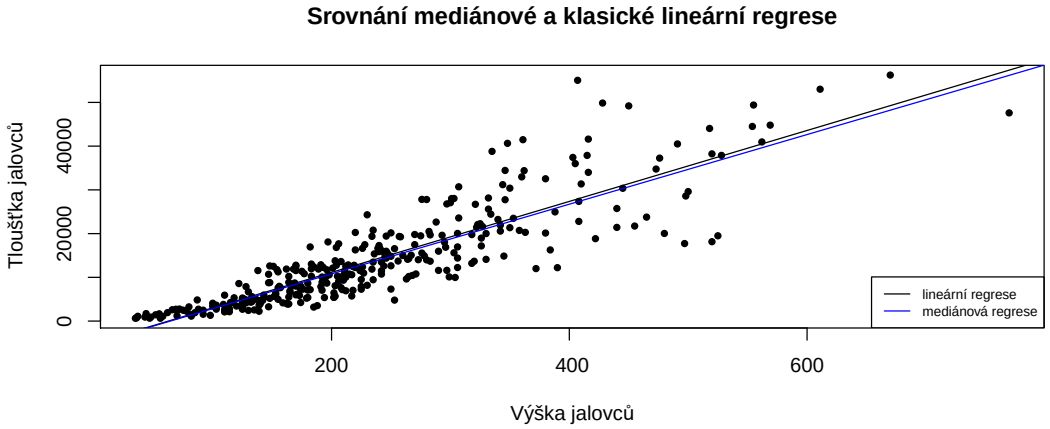
parametr	$\tau = 0,25$	$\tau = 0,5$	$\tau = 0,75$	lineární regrese
β_0	-4600,00	-5123,19	-4315,24	-5086,00
β_1	66,43	79,62	90,40	81,11

Tabulka 5.6: Hodnoty regresních parametrů kvantilové regrese pro různé kvantily a srovnání s lineární regresí

Parametr $\beta_1(\tau)$ můžeme interpretovat jako hodnotu, o kterou se změní podmíněný τ-kvantil proměnné tloušťka, změní-li se hodnota výšky o jednotku. Parametr $\beta_0(\tau)$ hraje důležitou roli z hlediska polohy regresní přímky. Samotnou hodnotu tohoto parametru však pro nás nemá smysl interpretovat, neboť by v podstatě vyjadřovala podmíněný τ-kvantil tloušťky jalovce při jeho nulové výšce.



Obrázek 5.6: Regresní přímky pro různé kvantily



Obrázek 5.7: Srovnání mediánové a klasické lineární regrese

Pokud jde o intervalové odhady vektoru regresních parametrů, je situace o něco komplikovanější. Máme totiž několik možností, jak tyto odhady konstruovat. Jednou možností je použití funkce

`summary.rq(object,se)`.

Nejdůležitějším parametrem je `object`, což je model kvantilové regrese, v našem případě tedy `model1`.

Parametr `se` je volitelný a značí metodu výpočtu směrodatné odchylky odhadů regresních parametrů.

Jednou z těchto metod je metoda *"rank"*. Tato metoda je brána jako defaultní, pokud je velikost souboru menší než 1000 pozorování, jako v našem případě. Výstupem funkce jsou při použití této metody kromě bodových odhadů regresních parametrů také hraniční hodnoty intervalů spolehlivosti těchto parametrů. Funkce tedy dává jako výstup samotné intervaly spolehlivosti. Konstrukce intervalových odhadů regresních parametrů je u této metody založena na neparametrickém pořadovém testu, který popisuje Koenker v [4]. Metoda zahrnuje řešení úlohy lineárního programování a pro velké výběry může být průběh výpočtu extrémně pomalý, proto se použití metody nedoporučuje pro soubory s velkým počtem pozorování. Defaultní nastavení navíc předpokládá, že jsou chybové členy nezávisle a stejně rozděleny. Tento předpoklad lze změnit použitím dodatečného argumentu funkce `summary.rq`, kterým je `iid = FALSE`.

Použijeme metodu *rank* pro získání intervalových odhadů regresních parametrů modelu (5.1)

```
summary.rq(model1,se="rank")
```

a dostaneme

```
Coefficients:
             coefficients lower bd    upper bd
(Intercept) -5123.18861  -6143.52504 -4143.13910
vyska          79.61566     72.58109    85.26114
Warning message:
In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) : Solution may be
nonunique.
```

Varovná hláška o nejednoznačnosti získaného řešení, která se objevuje ve výstupu funkce, je typická pro mediánovou regresi a ve stručnosti znamená, že při řešení úlohy lineárního programování v metodě *rank* simplexovou metodou, pokládá funkce za řešení již první z několika možných vrcholů. Tato situace nastává běžně v případě proměnných, které se svým charakterem blíží diskrétnímu typu.

Podobně je tato nejednoznačnost obvyklá pro výpočet mediánu v případě sudého počtu pozorování. U regrese pro jiné kvantily než medián se varovná hláška neobjevuje. Profesor Koenker odůvodňuje výskyt této hlášky v [6], kde také prohlašuje, že hláška nepředstavuje skutečný problém. Můžeme ji tedy dále ignorovat.

Hladinu α pro intervalový odhad specifikujeme pomocí parametru `alpha`. V defaultním nastavení je bráno `alpha = 0.1`, tedy je počítán 90% interval spolehlivosti. Chceme-li získat 95% interval, změníme argument na `alpha = 0.05`. Tedy

```
summary.rq(model1,se="rank",alpha=0.05)
```

a výstup se změní na

```
Coefficients:
             coefficients lower bd    upper bd
(Intercept) -5123.18861  -6143.52504 -4143.13910
vyska          79.61566     72.58109    85.26114
Warning message:
In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) : Solution may be
nonunique.
```

Získáváme tedy 95% intervaly spolehlivosti pro regresní parametry modelu (5.1):

$$\beta_0(0,5) \in (-6248,23035; -4068,23027)$$
$$\beta_1(0,5) \in (71,54646; 85,63816)$$

Z dalších metod jmenujme ještě metodu `boot`, která odhaduje směrodatnou odchylku odhadů regresních parametrů jednou z možných bootstrapových technik. Například bootstrapová metoda vybírání párů (x,y), která je popsána v podkapitole 4.2, se zvolí pomocí dodatečného argumentu `bsmethod = xy`. Výstupem funkce `summary.rq` za použití metody `boot` je tabulka bodových odhadů regresních parametrů, odhadů jejich směrodatných odchylek, t-statistik a p-hodnot, tedy stejně jako u výstupu funkce `summary` pro klasickou lineární regresi. Jelikož

jde však o bootstrapovou techniku, liší se hodnoty po každém použití funkce. Použijeme-li

```
summary.rq(model1,se="boot",bsmethod="xy"),
```

výstupy mají následující podobu:

```
Coefficients:
             Value       Std. Error  t value     Pr(>|t|)
(Intercept) -5123.18861   735.10962    -6.96929     0.00000
vyska          79.61566     4.07045    19.55940     0.00000.
```

Alternativně lze konstruovat intervaly spolehlivosti pomocí funkce `boot.rq`. Použitím této funkce vytvoříme model kvantilové regrese. Funkce má podobu

```
boot.rq(x,y,tau,R,bsmethod),
```

kde `x` je matice hodnot vysvětlujících proměnných, `y` je vektor hodnot vysvětlované proměnné, `tau` je požadovaný kvantil, pro který regresi provádíme, `R` je počet bootstrapových odhadů, tedy kolikrát bootstrapový odhad opakujeme. Konečně `bsmethod` je opět metoda, kterou pro bootstrapový odhad aplikujeme. Pro náš model (5.1) můžeme tedy funkci spustit kupříkladu takto:

```
model1<-boot.rq(cbind(1,vyska),tloustka,tau=0.5,R=10000,
bsmethod ="xy").
```

Intervalový odhad pro regresní parametry můžeme odtud získat percentilovou metodou popsanou v podkapitole 4.2 v rovnici (4.4). Například 95% interval získáme příkazem

```
t(apply(model1$B, 2, quantile, c(0.025,0.975))).
```

Výstupem jsou přímo intervalové odhady pro jednotlivé regresní parametry

```
              2.5%       97.5%
[1,] -6419.50954 -3577.77528
[2,]    70.00461    86.20588.
```

Můžeme si všimnout, jak se bootstrapový odhad liší oproti odhadu pomocí funkce `summary.rq` metodou `rank`. Bootstrap bude patrně výhodné použít v situacích, kdy budeme mít velké datové soubory.

Toto je jen několik možností, jak konstruovat intervalové odhady v programu R. Výčet mnoha dalších funkcí pro práci s modelem kvantilové regrese a podrobnosti týkající se jejich použití jsou k nalezení v [7].

V této práci budeme dále pro konstrukci intervalových odhadů regresních parametrů používat výhradně funkci `summary.rq` s metodou `rank` a hladinou `alpha = 0.05`.

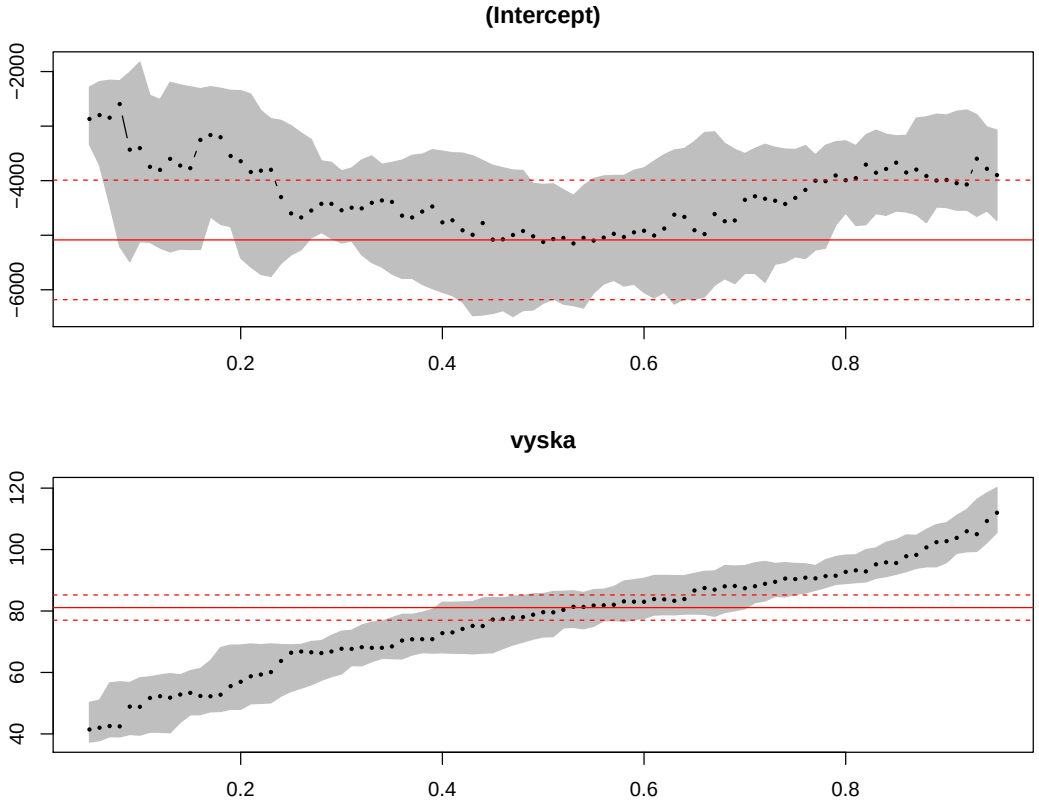
Jelikož hovoříme o kvantilové regresi, nezajímají nás intervalové odhady regresních parametrů pouze v mediánu, nýbrž napříč celým kvantilovým rozdělením. Tyto intervaly si můžeme graficky znázornit následujícím způsobem:

```
intervaly<-rq(tloustka~vyska,tau=seq(from=0.05,to=0.95,by=0.01))
plot(summary(intervaly,alpha=.05)).
```

Takto získáme obrázek 5.8. Obrázek obsahuje dva grafy. První graf pro parametr $\beta_0(\tau)$, druhý pro parametr $\beta_1(\tau)$. Černé tečky v grafech jsou bodové odhady těchto parametrů pro 0,05 až 0,95-kvantil s krokem 0,01. Horní a dolní mez i krok lze samozřejmě dle potřeby upravit v argumentech funkce. Šedý pás kolem bodových odhadů představuje 95% interval spolehlivosti daného parametru. Pro srovnání jsou v grafech červeně vyznačeny také odhady pro klasický lineární regresní model. Plná červená čára představuje bodový odhad parametru, přerušované červené čáry jsou pak hranice 95% intervalu spolehlivosti tohoto parametru.

Je vidět, že bodový i intervalový odhad parametru v případě klasické lineární regrese nezávisí na hodnotě kvantilu, jelikož je odhadována podmíněná střední hodnota. Naproti tomu, odhady parametrů kvantilové regrese se liší podle toho, jaký podmíněný kvantil je odhadován. Díky této vlastnosti může kvantilová regrese detekovat odlišnosti, které obyčejná lineární regrese nedokáže postřehnout. Můžeme zkoumat rozdíly v odhadech parametrů napříč kvantilovým rozdělením.

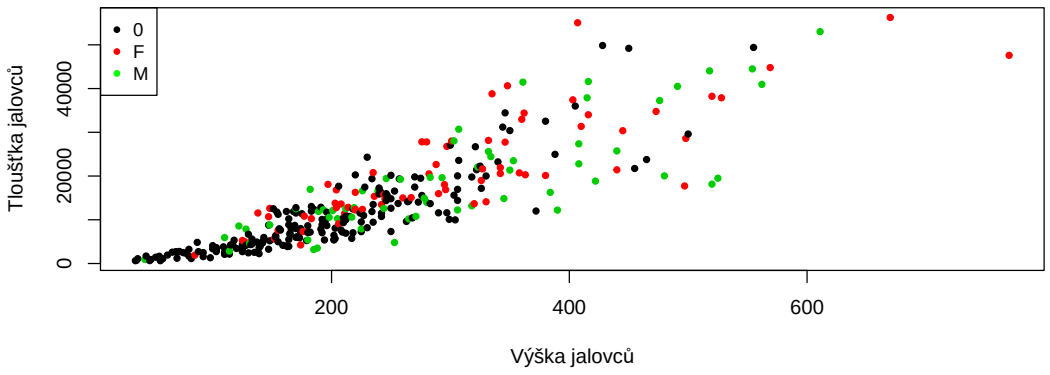
Například odhad parametru β_1 u lineární regrese se překrývá s odhadem tohoto parametru u kvantilové regrese v mediánu, avšak při změně odhadovaného kvantilu se odhady rozcházejí. V extrémech jako je odhadovaný podmíněný 0,05 nebo 0,95-kvantil jsou již odhady parametru zcela mimo pás odhadu lineární regrese. Znamená to tedy, že ačkoliv lineární regrese dokáže dobře vystihnout chování okolo mediánu, informace, kterou nám podává, nemá žádný význam pro velké či malé kvantily, a proto není pro popis regresního chování napříč celým kvantilovým rozdělením vhodná. Použití kvantilové regrese má tímto své odůvodnění.



Obrázek 5.8: Bodové odhady (černé tečky) a 95% intervaly spolehlivosti (šedou barvou) pro jednotlivé regresní parametry podle kvantilu v modelu 5.1. Červenou barvou je znázoněn bodový (plnou čarou) a 95% intervalový (přerušovanou čarou) odhad těchto parametrů pomocí klasické lineární regrese

5.3. Model kvantilové regrese s jednou kvantitativní a jednou kategoriální vysvětlující proměnnou

Model v podkapitole 5.2 byl záměrně zjednodušený, abychom si názorně ukázali základní postupy při vytváření a práci s modelem kvantilové regrese. Již v podkapitole 5.1 však vyšlo najevo, že by bylo vhodné rozlišit jalovce v souboru podle kategorie pohlaví. Toto rozdělení jsme v předchozí podkapitole vůbec nezohlednili. Obrázek 5.9 ukazuje bodový graf závislosti tloušťky na výšce tentokrát s barevným rozlišením jalovců podle kategorie pohlaví.



Obrázek 5.9: Bodový graf závislosti tloušťky na výšce s rozlišením pohlaví

Nebudeme již uvažovat přímkový model, ale pokusíme se data proložit parabolou. Jako kvantitativní vysvětlující proměnnou proto použijeme výšku v její druhé mocnině.

Druhou proměnnou je pohlaví. Do modelu zahrneme pouze interakce mezi proměnnými *ln(výška)* a *pohlaví*. Absolutní člen tedy zůstane pro všechny kategorie pohlaví stejný. Podle vzorce (1.27) můžeme tedy model zapsat takto:

$$\begin{aligned} tloušťka \;=\; & \beta_0(\tau) + \beta_1(\tau) \cdot výška^2 + \beta_2(\tau) I_{[F]} \cdot výška^2 + \\ & + \beta_3(\tau) I_{[M]} \cdot výška^2 + \epsilon(\tau). \end{aligned} \tag{5.2}$$

Předpis modelu zadáváme do funkce `rq` v podobě

```
model2<-rq(tloustka~I(vyska^2)+I(vyska^2):pohlavi,tau=...).
```

Pro každou kategorii pohlaví takto získáme odlišný regresní předpis. Kategorie neurčitého pohlaví je chápána jako výchozí, proto její předpis vypadá jednoduše:

$$neurčité : tloušťka = \beta_0(\tau) + \beta_1(\tau) \cdot výška^2 + \epsilon(\tau).$$

Předpis modelu pro jalovce samčího a samičího pohlaví obsahuje navíc proměnnou značící interakci mezi touto kategorií pohlaví a druhou mocninou výšky. Tedy pro samičí jalovce

$$samičí : tloušťka = \beta_0(\tau) + \beta_1(\tau) \cdot výška^2 + \beta_2(\tau) \cdot výška^2 + \epsilon(\tau)$$

a pro samčí jalovce

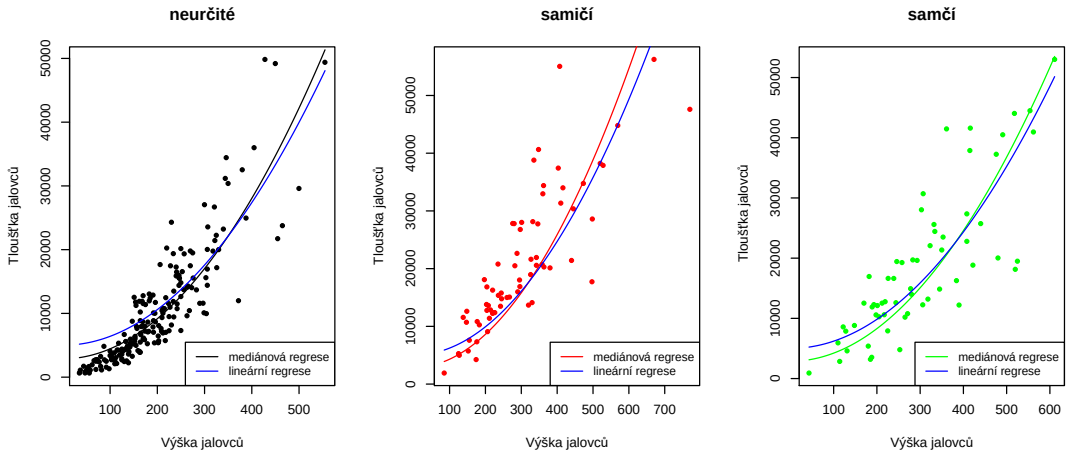
$$samčí : tloušťka = \beta_0(\tau) + \beta_1(\tau) \cdot výška^2 + \beta_3(\tau) \cdot výška^2 + \epsilon(\tau).$$

V tabulce 5.7 můžeme vidět bodové odhady jednotlivých parametrů modelu (5.2) pro vybrané kvantily a pro lineární regresní model se stejným předpisem. Z tabulky je patrné, že parametr β_2 vychází pro horní kvartil nulový, stejně tak parametr β_3 pro 0,95-kvantil. Regresní křivka pro jalovce neurčitého a samičího pohlaví v horním kvartilu je tak totožná a také je totožná regresní křivka pro jalovce samčího a neurčitého pohlaví v 0,95-kvantilu. Dále, záporné znaménko u hodnot odhadů parametrů β_2 a β_3 pro většinu kvantilů znamená, že regresní křivky pro samčí a samičí jalovce leží u těchto kvantilů pod křivkou pro jalovce neurčitého pohlaví.

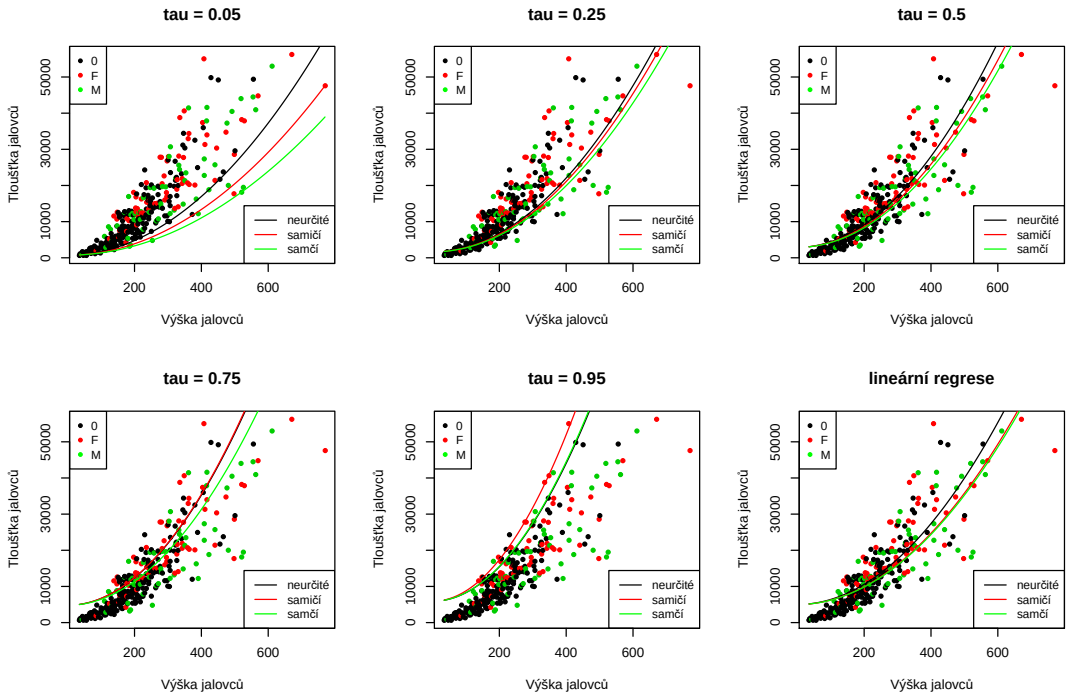
parametr	$\tau = 0{,}05$	$\tau = 0{,}25$	$\tau = 0{,}5$	$\tau = 0{,}75$	$\tau = 0{,}95$	lineární regrese
β_0	712,55	1622,54	2869,51	4784,76	5887,23	5003,35
β_1	0,10	0,13	0,16	0,19	0,24	0,14
β_2	-0,02	-0,01	-0,01	0,00	0,05	-0,02
β_3	-0,04	-0,01	-0,02	-0,02	0,00	-0,02

Tabulka 5.7: Bodové odhady regresních parametrů modelu (5.2) pro různé kvantily a pro lineární regresi. Odhady jsou zaokrouhleny na dvě desetinná místa.

Srovnání lineární a mediánové regrese pro jednotlivé kategorie pohlaví je graficky znázorněno na obrázku 5.10.



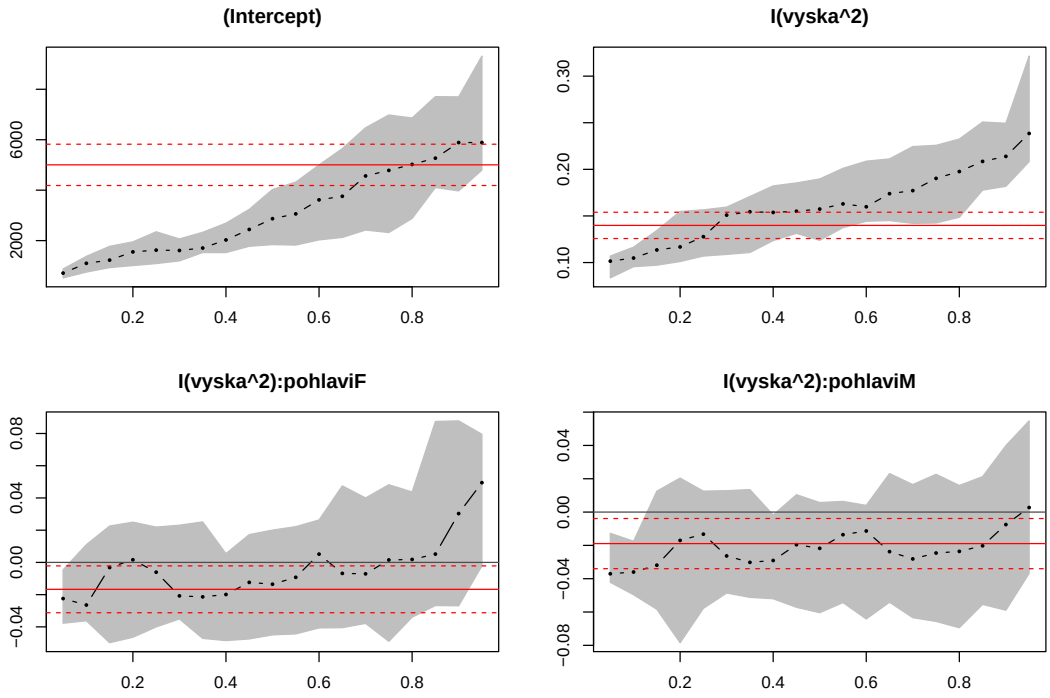
Obrázek 5.10: Srovnání lineární a mediánové regrese jednotlivé kategorie pohlaví v modelu (5.2)



Obrázek 5.11: Srovnání regresních křivek podle pohlaví pro různé kvantily a pro lineární regresi v modelu (5.2)

Na obrázku 5.11 jsou vykresleny pro srovnání regresní křivky podle kategorie pohlaví pro vybrané kvantily a lineární regresní model podle tabulky 5.7.

Abychom mohli posoudit, zda jsou rozdíly v regresních předpisech mezi pohlavími významné, potřebujeme více informací, než kolik poskytují bodové odhady. Z toho důvodu budeme konstruovat také intervalové odhady regresních parametrů napříč kvantilovým rozdělením.



Obrázek 5.12: 95% intervaly spolehlivosti regresních parametrů modelu (5.2). Nula je symbolizována plnou vodorovnou černou čarou.

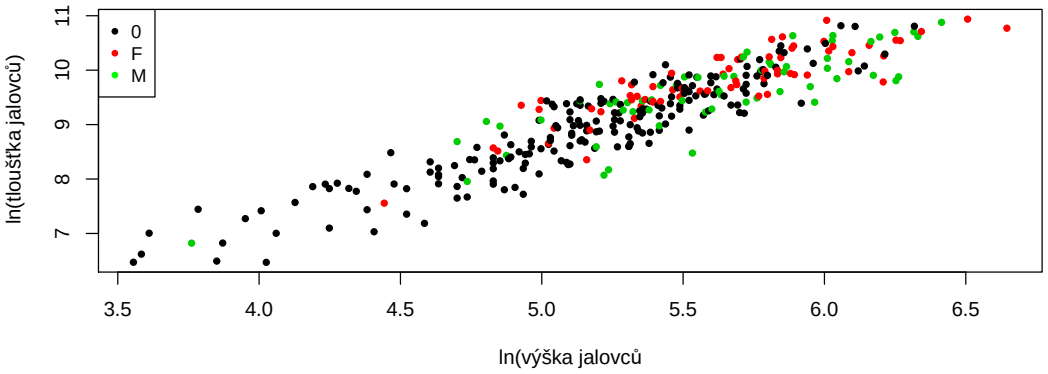
Intervalové odhady parametrů modelu (5.2) jsou vykresleny na obrázku 5.12 pro 0,05 až 0,95 kvantil s krokem 0,05.

Z obrázku je patrné, že pás spolehlivosti pro parametry β_2 a β_3 u většiny kvantilů zahrnuje nulu. To znamená, že jsou tyto parametry nevýznamné, tedy je statisticky nevýznamný rozdíl mezi regresními křivkami pro jalovce různého pohlaví. Model (5.2) tedy neodhalil rozdíly mezi pohlavími a proměnná *pohlaví* se proto jeví jako zbytečná.

Pokusíme se vytvořit ještě jiný model, tentokrát s využitím transformace

vysvětlované proměnné přirozeným logaritmem.

$$ln(tloušťka) = \beta_0(\tau) + \beta_1(\tau) \cdot ln(výška) + \beta_2(\tau) I_F \cdot ln(výška) + \\ + \beta_3(\tau) I_M \cdot ln(výška) + \epsilon(\tau). \tag{5.3}$$

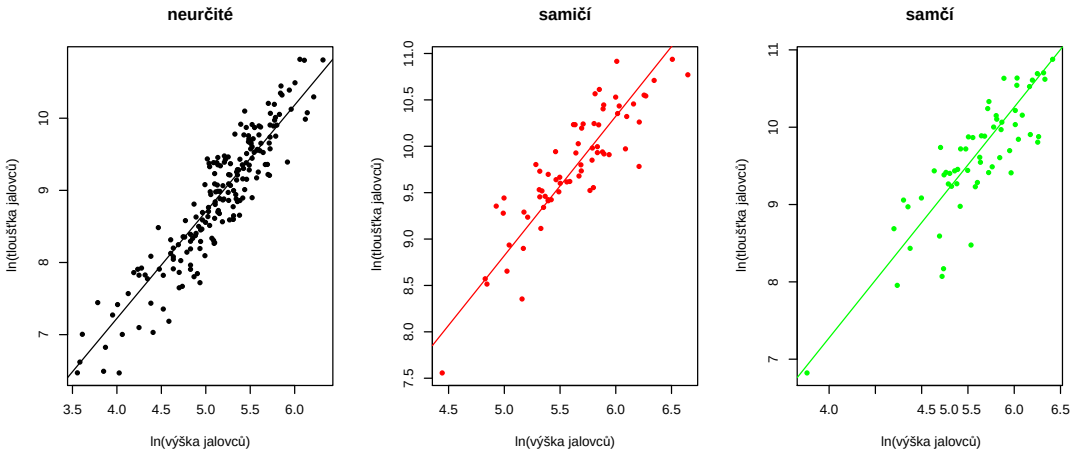


Obrázek 5.13: Bodový graf závislosti přirozeného logaritmu tloušťky na přirozeném logaritmu výšky s rozlišením pohlaví

Na obrázku 5.13 je bodový graf závislosti tloušťky na výšce po transformaci vysvětlované proměnné. Mediánová regrese pro tato data s rozlišením kategorie pohlaví je znázorněna graficky na obrázku 5.14.

Přestože je vhodné znázornit, jak dobře logaritmus prokládá transformovaná data, chtěli bychom si spíše ukázat, jak vypadají regresní předpisy a křivky modelu (5.3) pro původní data. V podkapitole 3.2 je uvedeno, že v případě kvantilové regrese získáme parametry původního modelu aplikací inverzní transformace na data transformovaného modelu, v našem případě použitím exponenciály na model transformovaný přirozeným logaritmem (viz rovnice (3.8) a (3.9))

Například regresní předpis pro odhadnutý model závislosti tloušťky na výšce u jalovců neurčitého pohlaví odvodíme takto:



Obrázek 5.14: Mediánová regrese pro transformovaná data v modelu (5.3) s rozlišením kategorie pohlaví

$$ln(\hat{tloušťka}) = \hat{\beta}_0(\tau) + \hat{\beta}_1(\tau) \cdot ln(výška),$$
$$e^{ln(\hat{tloušťka})} = e^{\hat{\beta}_0(\tau)+\hat{\beta}_1(\tau)\cdot ln(výška)},$$
$$\hat{tloušťka} = e^{\hat{\beta}_0(\tau)} \cdot výška^{\hat{\beta}_1(\tau)}$$

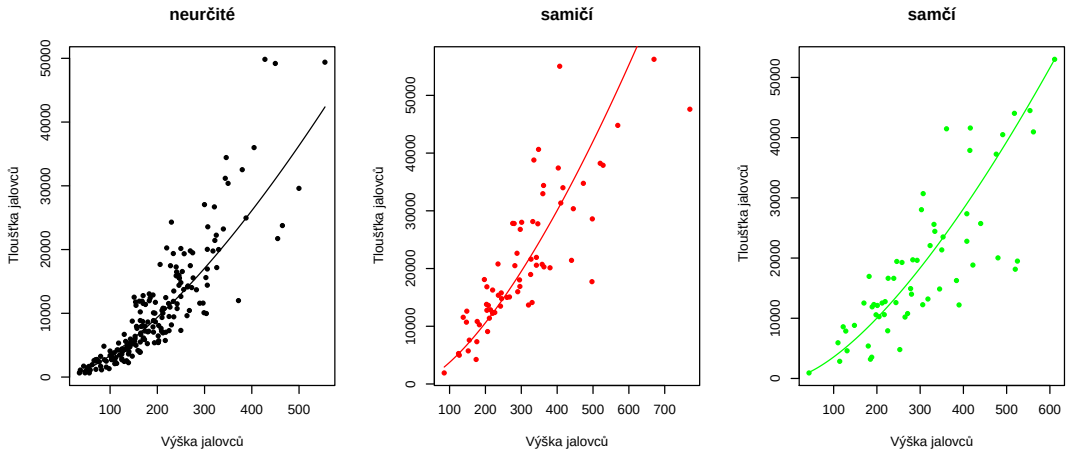
Obdobně bychom odvodili pro samičí a samčí jalovce

$$samičí: \hat{tloušťka} = e^{\hat{\beta}_0(\tau)} \cdot výška^{\hat{\beta}_1(\tau)+\hat{\beta}_2(\tau)},$$
$$samčí: \hat{tloušťka} = e^{\hat{\beta}_0(\tau)} \cdot výška^{\hat{\beta}_1(\tau)+\hat{\beta}_3(\tau)}.$$

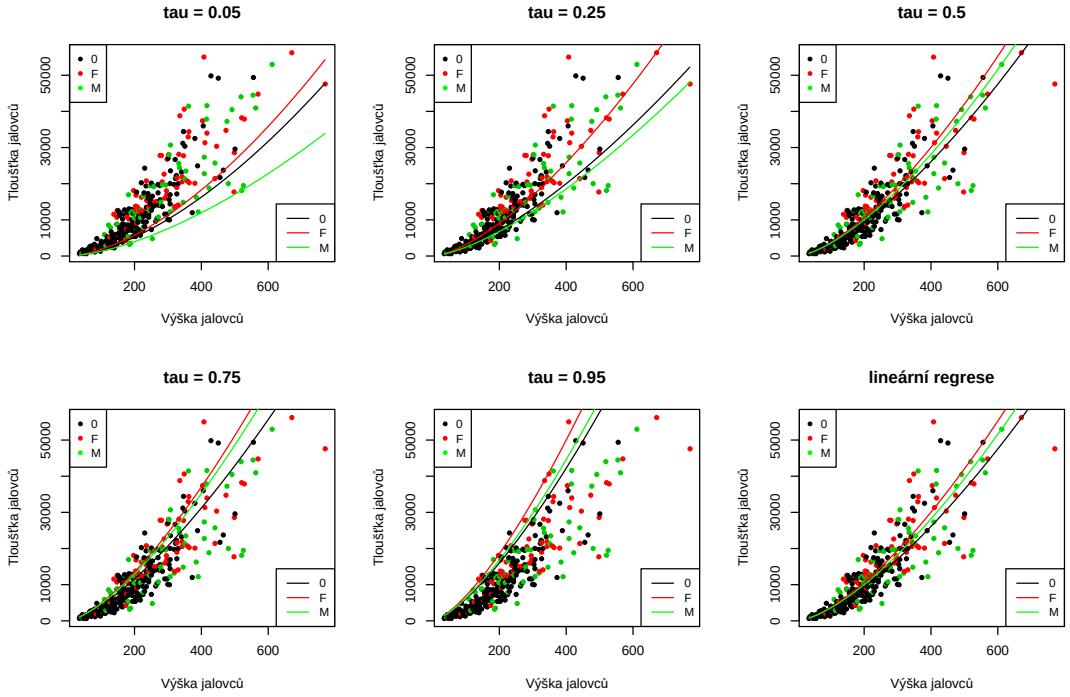
Celkově tedy předpis pro původní model dostaneme z předpisu pro model s transformací v podobě:

$$\hat{tloušťka} = e^{\hat{\beta}_0(\tau)} \cdot výška^{\hat{\beta}_1(\tau)+\hat{\beta}_2(\tau)\cdot I_{[F]}+\hat{\beta}_3(\tau)\cdot I_{[M]}}.$$

Tento předpis však již není lineární v parametrech. Odvozovali jsme jej pouze pro to, abychom znali předpisy regresních křivek pro grafické zobrazení, které je na obrázcích 5.15 a 5.16. Samotné hodnocení modelu budeme provádět na původním předpisu (5.3).



Obrázek 5.15: Mediánová regrese pro jednotlivé kategorie pohlaví v modelu (5.3)

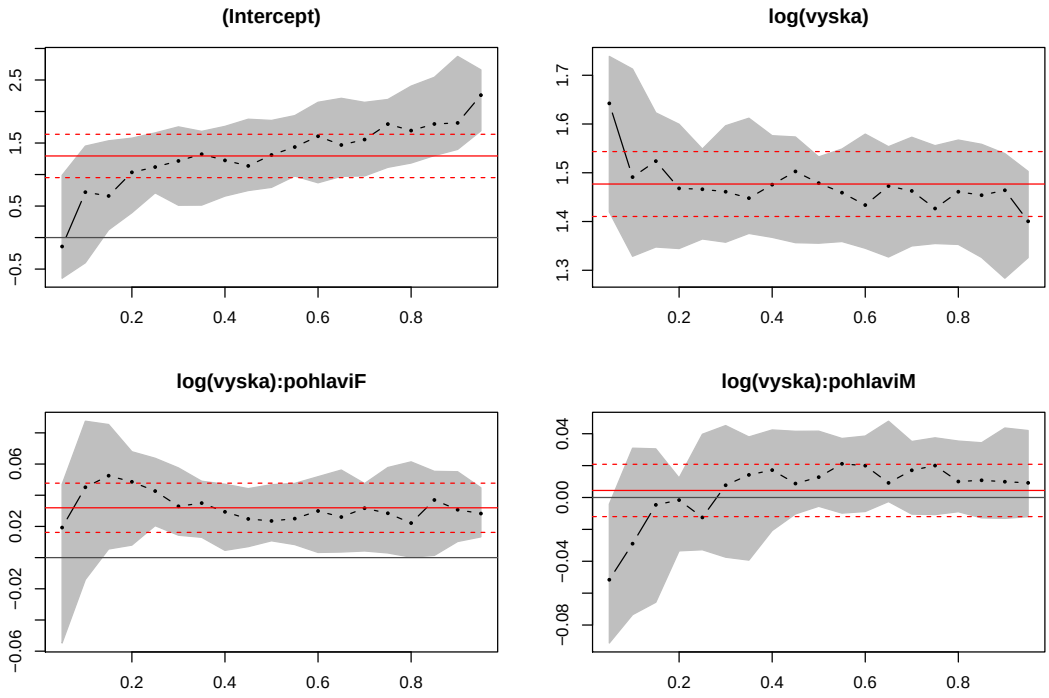


Obrázek 5.16: Srovnání regresních křivek u vybraných kvantilů a lineární regrese v modelu (5.3))

Na obrázku 5.16 vidíme, že u 0,05 kvantilu je křivka pro samčí jalovce položena nejníže, stejně jako v případě modelu (5.2). Až u větších kvantilů se křivka

pro samčí jalovce dostává nad křivku pro jalovce neurčitého pohlaví, zatímco křivka pro samičí jalovce je všude položena nejvýše.

Nyní se podíváme na intervalové odhady modelu (5.3). Tyto intervaly jsou graficky znázorněny na obrázku 5.17 pro 0,05 až 0,95-kvantil s krokem 0,05. Parametr $\beta_3(\tau)$ vychází opět pro téměř všechny kvantily nevýznamný, zato parametr β_2 je kromě velmi malých kvantilů. Z modelu tedy můžeme usuzovat, že u většiny kvantilů se regresní křivka pro samičí jalovce liší od regresních křivek pro zbylé dvě kategorie (je položena výše), zatímco mezi kategoriemi jalovců samčího a neurčitého pohlaví nejsou prokazatelné rozdíly v regresní závislosti.



Obrázek 5.17: 95% intervaly spolehlivosti regresních parametrů modelu (5.3)

Na závěr ještě srovnáme modely (5.2) a (5.3) pomocí pseudokoeficientu determinace. Hodnoty pseudokoeficientu nejsou ovlivněny změnou počtu parametrů modelu, neboť oba modely mají tyto parametry právě čtyři.

Tabulka 5.8 zobrazuje hodnoty pseudokoeficientu determinace modelů (5.2) a (5.3) pro vybrané kvantily počínaje 0,05 kvantilem až po 0,95 kvantil s krokem

0,05. V posledním řádku jsou ještě tyto hodnoty napříč všemi těmito kvantily zprůměrovány.

τ	model (5.2)	model (5.3)
0,05	0,388	0,665
0,10	0,387	0,638
0,15	0,401	0,633
0,20	0,421	0,631
0,25	0,439	0,628
0,30	0,459	0,625
0,35	0,473	0,622
0,40	0,480	0,618
0,45	0,485	0,611
0,50	0,489	0,605
0,55	0,497	0,599
0,60	0,508	0,594
0,65	0,520	0,590
0,70	0,528	0,586
0,75	0,542	0,582
0,80	0,562	0,581
0,85	0,587	0,584
0,90	0,609	0,586
0,95	0,641	0,587
průměr	0,495	0,609

Tabulka 5.8: Srovnání pseudokoeficientu determinace modelu (5.2) a (5.3) pro vybrané kvantily

Z tabulky je zřejmé, že z hlediska pseudokoeficientu determinace celkově vychází jako lepší model 5.3. Pouze pro některé velké kvantily vychází hodnoty pseudokoeficientu vyšší u modelu 5.2.

5.4. Model kvantilové regrese s dvěma kvantitativními a jednou kategoriální proměnnou

V této podkapitole model kvantilové regrese ještě o něco zkomplikujeme přidáním druhé kvantitativní proměnné, kterou je přirozený logaritmus nadmořské výšky.

Budeme stejně jako v podkapitole 5.3 uvažovat dva modely, jeden s původním vyjádřením vysvětlované proměnné, druhý za použití logaritmické transformace. Tentokrát ponecháme pravou stranu předpisu stejnou pro oba modely, abychom

mohli pozorovat, jaký vliv má pouze samotná logaritmická transformace na výstupy modelu.

Nejprve se podíváme na model bez transformace vysvětlované proměnné. Jeho předpis je vyjádřen rovnicí

$$\begin{aligned} tloušťka =& \beta_0(\tau) + \beta_1(\tau) \cdot ln(výška) + \beta_2(\tau) \cdot I_{[F]} \cdot ln(výška) + \\ &+ \beta_3(\tau) \cdot I_{[M]} \cdot ln(výška) + \beta_4(\tau) \cdot ln(nadmořská\ výška) + \epsilon(\tau). \end{aligned} \tag{5.4}$$

Tabulka 5.9 ukazuje bodové odhady jednotlivých regresních parametrů modelu (5.4). Můžeme si povšimnout, že odhady parametru β_4 jsou záporné, nadmořská výška tedy působí na tloušťku jalovců negativně (s rostoucí nadmořskou výškou mají jalovce tendenci být tenčí). To by odpovídalo intuitivní představě, že ve vyšších polohách jsou díky nízké teplotě a malému vegetativnímu porostu pro růst jalovců nepříznivé podmínky.

parametr	$\tau = 0{,}05$	$\tau = 0{,}25$	$\tau = 0{,}5$	$\tau = 0{,}75$	$\tau = 0{,}95$	lineární regrese
β_0	38075,51	126813,37	175395,27	180979,04	626219,50	240833,63
β_1	9414,50	10944,65	13125,50	14630,98	14775,57	14924,49
β_2	492,23	588,55	460,41	1116,28	956,86	635,23
β_3	102,47	266,39	403,56	307,14	888,55	302,87
β_4	-11371,74	-24453,25	-32444,72	-33814,99	-94355,23	-42639,21

Tabulka 5.9: Bodové odhady regresních parametrů modelu (5.4) pro různé kvantily a pro lineární regresi. Odhady jsou zaokrouhleny na dvě desetinná místa.

Intervaly spolehlivosti parametrů modelu (5.4) jsou ovšem příliš široké. Pro volbu kvantilu $\tau = 0,5$ vycházejí tyto intervaly:

```
Coefficients:
                      coefficients lower bd     upper bd
(Intercept)          175395.27370  93839.85379 337922.39185
log(vyska)            13125.50352  10485.95980  15210.32972
log(nadmorska_vyska) -32444.72379 -63777.31330 -20167.44893
log(vyska):pohlaviF     460.41033    145.41513   1014.58508
log(vyska):pohlaviM     403.56242    -55.22155    848.10522.
```

Při volbě některých jiných kvantilů jsou již vypočítané intervalové odhady tak široké, že je software R ani nedokáže graficky zobrazit. Například pro kvantil $\tau = 0,05$:

```
Coefficients:
                     coefficients  lower bd      upper bd
(Intercept)           3.807551e+04 -1.297953e+05  7.126649e+05
log(vyska)            9.414502e+03  6.721371e+03  1.545495e+04
log(nadmorska_vyska) -1.137174e+04 -4.660325e+04  1.797280e+04
log(vyska):pohlaviF   4.922259e+02  2.388270e+02  7.048459e+02
log(vyska):pohlaviM   1.024684e+02 -2.085329e+14  2.918504e+02.
```

Odtud vyplývá, že model (5.4) není vhodný. Namísto něj budeme uvažovat model (5.5), kdy transformací vysvětlované proměnné přirozeným logaritmem zúžíme intervaly spolehlivosti.

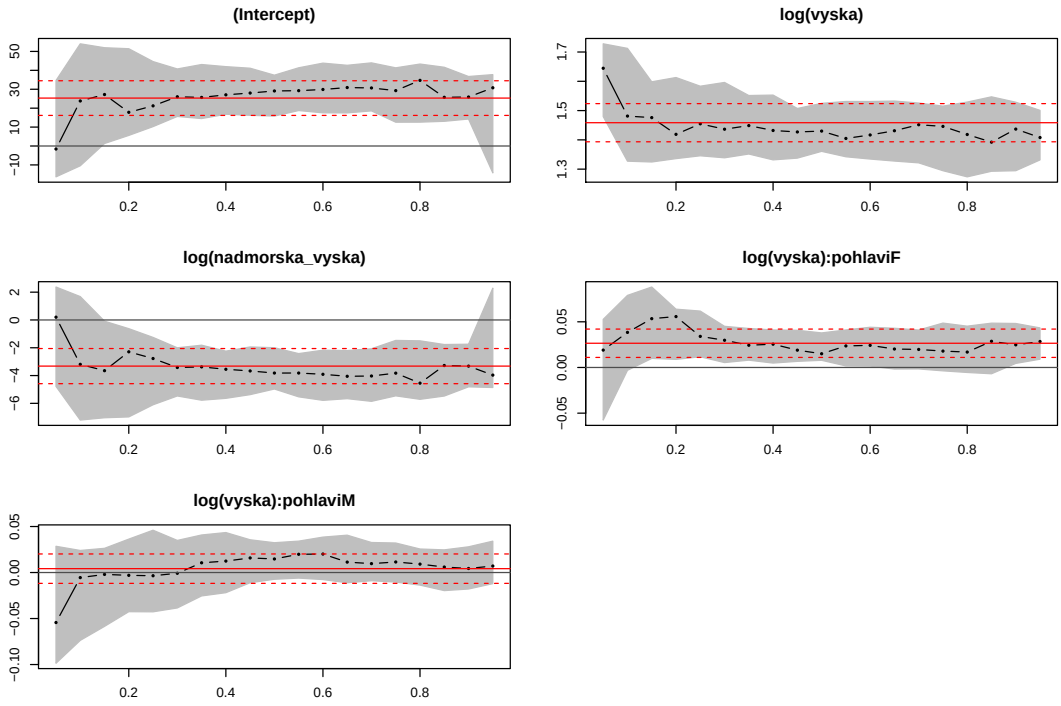
$$\begin{aligned} ln(tloušťka) =& \beta_0(\tau) + \beta_1(\tau) \cdot ln(výška) + \beta_2(\tau) \cdot I_{[F]} \cdot ln(výška) + \\ &+ \beta_3(\tau) \cdot I_{[M]} \cdot ln(výška) + \beta_4(\tau) \cdot ln(nadmořská\ výška) + \epsilon(\tau). \end{aligned} \tag{5.5}$$

Pro odhad podmíněného mediánu v modelu (5.5) již skutečně dostaneme:

```
Coefficients:
                     coefficients lower bd upper bd
(Intercept)          29.09881     15.82315 37.44499
log(vyska)            1.42997      1.36077  1.52467
log(nadmorska_vyska) -3.81899     -4.96857 -1.99848
log(vyska):pohlaviF   0.01503      0.00786  0.03784
log(vyska):pohlaviM   0.01475     -0.00734  0.03239.
```

Dále znázorníme intervaly spolehlivosti modelu (5.5) graficky a obdobně jako

v předchozí podkapitole srovnáme pseudokoeficienty determinace modelů (5.4) a (5.5). Můžeme na základě tabulky 5.10 vidět, že logaritmická transformace pseudokoeficient determinace výrazně zvýšila.



Obrázek 5.18: 95% intervaly spolehlivosti regresních parametrů modelu (5.5)

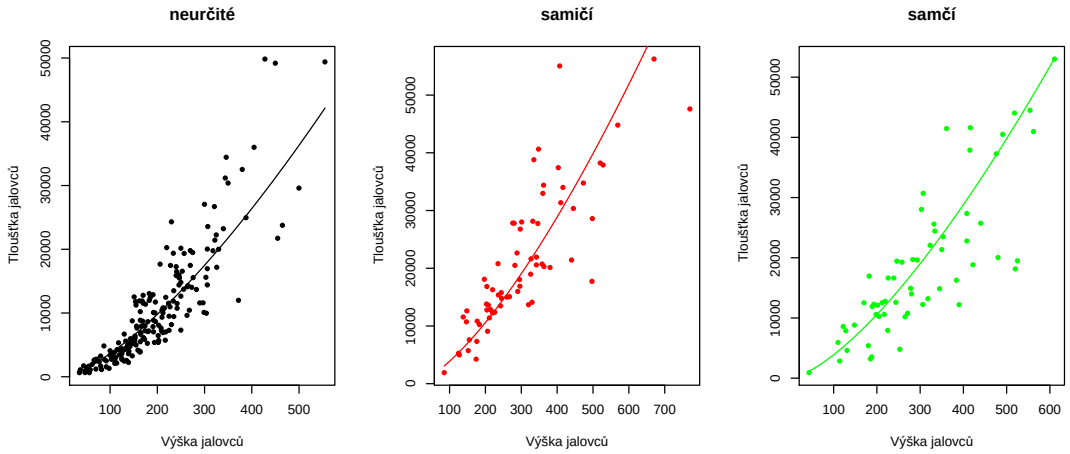
Můžeme vyjádřit předpis pro model bez transformace vysvětlované proměnné inverzní transformací, stejně jako v podkapitole 5.3

$$\widehat{tloušťka} = e^{\hat{\beta}_0(\tau)} \cdot výška^{\hat{\beta}_1(\tau)+\hat{\beta}_2(\tau)\cdot I_{[F]}+\hat{\beta}_3(\tau)\cdot I_{[M]}} \cdot nadmořská\ výška^{\hat{\beta}_4(\tau)}.$$

Tento předpis je opět nelineární, umožňuje však graficky vykreslit regresní křivky. Jelikož máme tentokrát dvě kvantitativní vysvětlující proměnné, abychom mohli zobrazit regresní křivky do dvojrozměrného grafu, stanovíme pevnou hodnotu nadmořské výšky jako její medián, což je 1335, 54 m.n.m. (viz tabulka 5.1), a výška zůstane jako proměnná.

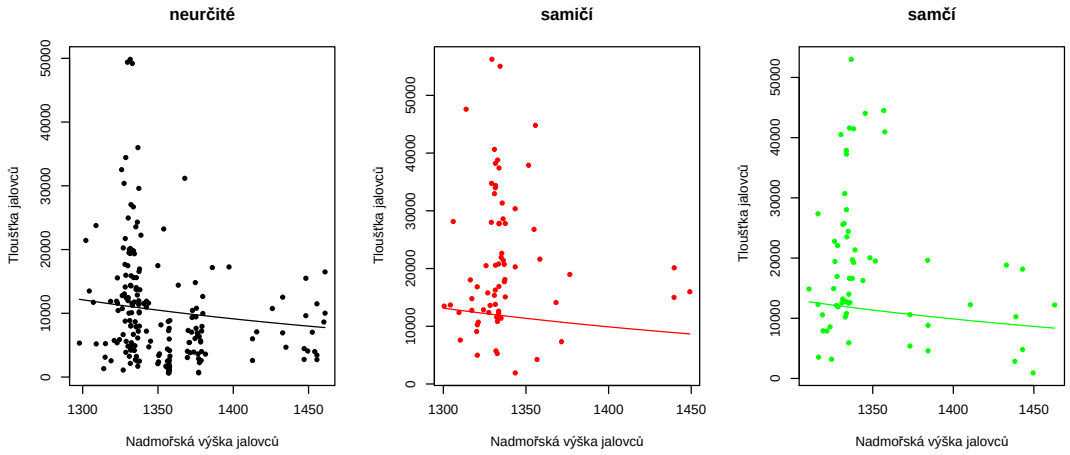
τ	model (5.4)	model (5.5)
0,05	0,374	0,665
0,10	0,392	0,642
0,15	0,411	0,639
0,20	0,426	0,637
0,25	0,436	0,638
0,30	0,444	0,639
0,35	0,448	0,636
0,40	0,451	0,632
0,45	0,447	0,627
0,50	0,443	0,621
0,55	0,442	0,616
0,60	0,445	0,611
0,65	0,446	0,606
0,70	0,446	0,602
0,75	0,444	0,598
0,80	0,447	0,596
0,85	0,460	0,601
0,90	0,463	0,604
0,95	0,459	0,603
průměr	0,438	0,622

Tabulka 5.10: Srovnání pseudokoeficientu determinace modelů (5.4) a (5.5) pro vybrané kvantily. Hodnoty jsou zaokrouhleny na tři desetinná místa



Obrázek 5.19: Mediánová regrese pro jednotlivé kategorie pohlaví v modelu (5.5) se zafixovanou hodnotou nadmořské výšky na jejím mediánu (= 1335, 54 m.n.m.).

Obrázek 5.20 zobrazuje pro srovnání regresní křivky pro mediánovou regresi v rámci jednotlivých kategorií pohlaví pro zafixovanou hodnotu výšky. Výška je pevně stanovena na hodnotě svého mediánu, tedy 216 cm.



Obrázek 5.20: Mediánová regrese pro jednotlivé kategorie pohlaví v modelu (5.5) se zafixovanou hodnotou výšky na jejím mediánu (= 216 cm).

Mírně klesající tvar regresních křivek na obrázku 5.20 poukazuje na negativní vliv nadmořské výšky na celkovou tloušťku jalovců.

Závěr

Záměrem bakalářské práce bylo seznámení se s hlavní myšlenkou kvantilové regrese, vytyčení podobností a rozdílů ve srovnání s klasickou lineární regresí a uplatnění získaných poznatků při analýze regresní závislosti mezi vybranými vlastnostmi jalovce obecného nízkého.

Teoretická část skládající se z prvních čtyř kapitol podává výklad o teoretických aspektech klasického lineárního regresního modelu a modelu kvantilové regrese. Dále tyto dva typy srovnává a ukazuje na analogie i zásadní rozdíly. Praktická část obsažená v páté kapitole se pak zaměřuje na konkrétní využití těchto znalostí.

Kvantilová regrese je oproti lineární regresi robustní vůči odlehlým pozorováním, nepředpokládá konkrétní parametrické rozdělení a zachovává způsob interpretace výstupu při monotónní transformaci vysvětlované proměnné. Dále je kvantilová regrese schopna modelovat celé podmíněné kvantilové rozdělení a sledovat tak vývoj regresní závislosti v okolí extrémních hodnot vysvětlované proměnné.

Naproti tomu je odhad parametrů ve srovnání s lineární regresí výpočetně náročnější a možnost použití asymptotického přístupu pro statistickou inferenci je silně limitována předpoklady, které jsou v praxi zřídka dodrženy. Díky zavedení výpočetní techniky, která zaznamenala v posledních několika desetiletích rychlý vývoj, však již výpočet odhadů regresních parametrů nepředstavuje problém a boostrapová metoda, kterou lze ve vhodném softwaru snadno implementovat, představuje vítanou alternativu k asymptotickému postupu pro inferenční statistiku.

Z těchto důvodů představuje kvantilová regrese užitečný nástroj v celé řadě

aplikací, jako jsou například risk management, medicína, ekologie, machine learning nebo společenské vědy. Více o kvantilové regresi a možnostech jejího uplatnění lze nalézt mimo jiné v [2] a [5].

Literatura

[1] FIŠEROVÁ, Eva. *Lineární statistické modely*. Olomouc, 2015. Skripta. Univerzita Palackého v Olomouci, Přírodovědecká fakulta.

[2] HAO, Lingxin a Daniel Q. NAIMAN. *Quantile Regression* [online]. London: Sage Publications, 2007. Dostupné také z: https://nguyenvantien0405.files.wordpress.com/2014/03/quantile_regressiolingxin-hao.pdf

[3] HRON, Karel, Pavla KUNDEROVÁ a Ondřej VENCÁLEK. *Základy počtu pravděpodobnosti a metod matematické statistiky*. Olomouc, 2018. Skripta. Univerzita Palackého v Olomouci, Přírodovědecká fakulta.

[4] KOENKER, Roger. *Confidence intervals for regression quantiles*. New York: Springer-Verlag, 1994.

[5] KOENKER, Roger. *Quantile Regression*. New York: Cambridge University Press, 2005.

[6] KOENKER, Roger. [R] Warning Messages using rq -quantile regressions. In: *R-help – Main R Mailing List: Primary help* [online]. 2006 [cit. 2022-03.-09]. Dostupné z: https://stat.ethz.ch/pipermail/r-help/2006-July/109821.html

[7] KOENKER, Roger. *Quantreg: Quantile Regression. R package version 5.85* [online]. 2021. Dostupné také z: https://CRAN.R-project.org/package=quantreg

[8] PROCHÁZKA, Jiří. *Kvantilová regrese* [online]. Praha, 2015. Dostupné také z: https://insis.vse.cz/zp/51570. Diplomová práce. Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky.

[9] RODRIGUEZ, Robert N. a Yonggang YAO. *Five Things You Should Know about Quantile Regression* [online]. 2017. Dostupné také z: https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf. SAS Institute Inc.

[10] ŠMÍDEK, Jakub. *Kvantilová regrese* [online]. Brno, 2008. Dostupné také z: https://is.muni.cz/th/a8pd1/Kvantilova_regrese.pdf. Bakalářská práce. Masarykova univerzita, Přírodovědecká fakulta.

[11] ZEIDLER, Miroslav, Eva FIŠEROVÁ, Veronika ŘÍMALOVÁ, Marek BEDNÁŘ a Marek BANAŠ. Sex-based differences in vigor and site preferences of Juniperus communis subsp. nana. *Nordic Journal of Botany* [online]. 2020, **2020**(38), 10. Dostupné z: https://onlinelibrary.wiley.com/doi/10.1111/njb.02812

Příloha A

Ukázkový kód v R

```
#nahrání balíčku Quantreg
library(quantreg)

#nahrání dat

#označení proměnných
#výška
vyska<-Data$height17

#tloustka - proměnná vypočítaná jako součin délky a šířky
Data$tloustka<-Data$lenght17*Data$width17
tloustka<-Data$tloustka

#pohlaví
levels(Data$sex17)
Data$sex17<-as.character(Data$sex17)
Data$sex17[Data$sex17=="F "]="F"
Data$sex17<-as.factor(Data$sex17)
levels(Data$sex17)
pohlavi<-Data$sex17
levels(pohlavi)

#nadmořská výška
nadmorska_vyska<-Data$altitude

#bodový graf - závislost tloušťky na výšce
plot(vyska,tloustka,pch=20,ylab="Tloušťka jalovců",xlab="Výška jalovců")

#vytvoření modelu kvantilové regrese
model1<-rq(tloustka~vyska,tau=0.5)

#bodové odhady parametrů
model1$coefficients

#intervalové odhady parametrů
summary.rq(model1,se="rank",alpha=0.05)

#grafické vykreslení mediánové regresní křivky
abline(model1,col="blue")

#můžeme vykreslit regresní křivku pro libovolný jiný kvantil, např. 0,9-kvantil
abline(rq(tloustka~vyska,tau=0.9),col="orange")

#graficky zobrazíme intervaly spolehlivosti pro jednotlivé regresní parametry
intervaly<-rq(tloustka~vyska,tau=seq(from=0.05,to=0.95,by=0.05))
plot(summary(intervaly,alpha=0.05))

#výpočet pseudokoeficientu determinace (musíme zadat ručně)
rho <- function(u,tau=0.5)u*(tau - (u < 0))
q<-seq(from=0.05,to=0.95,by=0.05)

R1<-c()
for (i in q){
        R1[i*20]<-1-rq(tloustka~vyska,tau=i)$rho/rq(tloustka~1,tau=i)$rho
}
R1
mean(R1)

#bodový graf závislosti tloušťky na výšce s rozlišením kategorie pohlaví
plot(vyska,tloustka,col=pohlavi,pch=20,xlab="Výška jalovců",ylab="Tloušťka jalovců")
legend("topleft",legend=c("0","F","M"),pch=c(20,20,20),col=c("black","red","green"),cex=0.6)
```

```
#model (5.2)
model2<-rq(tloustka~I(vyska^2)+I(vyska^2):pohlavi,tau=0.5)

#graficky vykreslíme mediánovou regresi pro jednotlivé kategorie pohlaví
x<-vyska

plot(vyska[pohlavi=="0"],tloustka[pohlavi=="0"],pch=20,xlab="Výška jalovců",ylab="Tloušťka jalovců",main = "neurčité")
curve(model2$coefficients[1]+model2$coefficients[2]*x^2,add=TRUE)
plot(vyska[pohlavi=="F"],tloustka[pohlavi=="F"],col="red",pch=20,ylab="Tloušťka jalovců",xlab="Výška jalovců",main = "samičí")
curve(model2$coefficients[1]+model2$coefficients[2]*x^2+model2$coefficients[3]*x^2,col="red",add=TRUE)
plot(vyska[pohlavi=="M"],tloustka[pohlavi=="M"],col="green",pch=20,ylab="Tloušťka jalovců",xlab="Výška jalovců",main="samčí")
curve(model2$coefficients[1]+model2$coefficients[2]*x^2+model2$coefficients[4]*x^2,col="green",add=TRUE)

#intervaly spolehlivosti
intervaly2<-rq(tloustka~I(vyska^2):pohlavi+I(vyska^2),tau=seq(from=0.05,to=0.95,by=0.05))
plot(summary(intervaly2,alpha=0.05))

#pseudokoeficient determinace
R2<-c()
for (i in q){
        R2[i*20]<-1-rq(tloustka~I(vyska^2):pohlavi+I(vyska^2),tau=i)$rho/rq(tloustka~1,tau=i)$rho
}
R2
mean(R2)

#bodový graf závislosti po logaritmické transformaci
plot(log(vyska),log(tloustka),col=pohlavi,pch=20,ylab="ln(tloušťka jalovců)",xlab="ln(výška jalovců")
legend("topleft",legend=c("0","F","M"),pch=c(20,20,20),col=c("black","red","green"),cex=0.6)

#model (5.3)
model3<-rq(log(tloustka)~log(vyska)+log(vyska):pohlavi,tau=0.5)

#grafické vykreslení mediánové regrese do transformovaných dat
plot(log(vyska)[pohlavi=="0"],log(tloustka)[pohlavi=="0"],pch=20,ylab="ln(tloušťka jalovců)",xlab="ln(výška jalovců)",main="neurčité")
abline(a=model3$coefficients[1],b=model3$coefficients[2],add=TRUE)
plot(log(vyska)[pohlavi=="F"],log(tloustka)[pohlavi=="F"],pch=20,col="red",ylab="ln(tloušťka jalovců)",xlab="ln(výška jalovců)",main="samičí")
abline(a=model3$coefficients[1],b=model3$coefficients[2]+model3$coefficients[3],col="red")
plot(log(vyska)[pohlavi=="M"],log(tloustka)[pohlavi=="M"],pch=20,col="green",ylab="ln(tloušťka jalovců)",xlab="ln(výška jalovců)",main="samčí")
abline(a=model3$coefficients[1],b=model3$coefficients[2]+model3$coefficients[4],col="green")

#grafické vykreslení mediánové regrese do původních dat
plot(vyska[pohlavi=="0"],tloustka[pohlavi=="0"],pch=20,xlab="Výška jalovců",ylab="Tloušťka jalovců",main = "neurčité")
curve(exp(model3$coefficients[1])*x^model3$coefficients[2],add=TRUE)
plot(vyska[pohlavi=="F"],tloustka[pohlavi=="F"],pch=20,col="red",xlab="Výška jalovců",ylab="Tloušťka jalovců",main = "samičí")
curve(exp(model3$coefficients[1])*x^(model3$coefficients[2]+model3$coefficients[3]),col="red",add=TRUE)
plot(vyska[pohlavi=="M"],tloustka[pohlavi=="M"],pch=20,col="green",xlab="Výška jalovců",ylab="Tloušťka jalovců",main = "samčí")
curve(exp(model3$coefficients[1])*x^(model3$coefficients[2]+model3$coefficients[4]),col="green",add=TRUE)

#intervaly spolehlivosti
intervaly3<-rq(log(tloustka)~log(vyska)+log(vyska):pohlavi,tau=seq(from=0.05,to=0.95,by=0.05))
plot(summary(intervaly3,alpha=0.05))

#pseudokoeficient determinace
R3<-c()
for (i in q){
        R3[i*20]<-1-rq(log(tloustka)~log(vyska)+log(vyska):pohlavi,tau=i)$rho/rq(log(tloustka)~1,tau=i)$rho
}
R3
mean(R3)
```