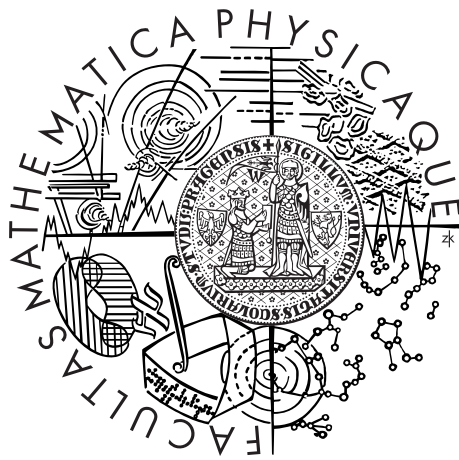


UNIVERZITA KARLOVA V PRAZE
MATEMATICKO-FYZIKÁLNÍ FAKULTA

DIPLOMOVÁ PRÁCE



Blanka Hamplová

Testování normality v lineárním modelu

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Karel Zvára, CSc.
Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán: Matematická statistika

Děkuji vedoucímu mé diplomové práce doc. RNDr. Karlu Zvárovi, CSc., za cenné rady a připomínky, které mi pomohly při vytváření této práce, a také za trpělivost a ochotu, se kterou si vždy našel čas na konzultaci. Dále děkuji všem, kteří moji práci přečetli a upozornili mne na řadu překlepů i věcných chyb.

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

V Praze dne

Blanka Hamplová

Obsah

Obsah	4
Abstrakt	5
Úvod	6
1 Lineární model	7
1.1 Normální lineární model	9
1.2 Další rezidua	9
1.2.1 Normovaná rezidua	9
1.2.2 Studentizovaná rezidua	10
1.2.3 Nekorelovaná rezidua	11
2 Testy normality	13
2.1 Shapirův-Wilkův test W	13
2.2 Shapirův-Francův test W'	14
2.3 Test založený na šikmosti ($\sqrt{b_1}$)	15
2.4 Test založený na špičatosti (b_2)	16
2.5 Souhrnný test K^2	17
2.6 Test T_3 založený na EMVF	17
2.7 Lillieforsův test D^*	19
2.8 Michaelův test D_{SP}	20
2.9 Cramerův-von Misesův test W^{2*}	21
2.10 Andersonův-Darlingův test A^{2*}	22
3 Rozdělení	23
3.1 Normální rozdělení	23
3.2 Rovnoměrné rozdělení	24
3.3 Dvojitě exponenciální rozdělení	24
3.4 Exponenciální rozdělení	25
3.5 Studentovo t -rozdělení	25
3.6 Smíšené rozdělení	26
4 Simulace	27
4.1 Ověření hladiny testů	27
4.2 Srovnání síly testů	31
4.2.1 Model regresní přímky	33
4.2.2 Model kvadratické závislosti	35
4.2.3 Model regrese se třemi regresory	37
4.2.4 Model jednoduchého třídění	38

Závěr	39
A Obrázky	40
Literatura	64

Abstrakt

Název práce: Testování normality v lineárním modelu

Autor: Blanka Hamplová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Karel Zvára, CSc.

e-mail vedoucího: karel.zvara@mff.cuni.cz

Abstrakt: Tato diplomová práce se zabývá způsoby ověřování předpokladu, že náhodná složka vysvětlované proměnné v lineárním modelu má normální rozdělení. V rámci simulační studie jsme sledovali jak použití různých typů reziduí místo původního chybového vektoru ovlivňuje dodržování hladiny deseti testů normality. Protože jednotlivá rezidua závisí na matici modelu, uvažovali jsme čtyři konkrétní matice běžně používaných modelů. Dále jsme srovnávali vliv použitého modelu a typu reziduí na sílu jednotlivých testů proti pěti různým alternativním rozdělením. Ukázalo se, že chování testů je spíše než modelem ovlivňováno použitým typem reziduí. Speciálně studentizovaná rezidua se jeví jako nevhodná, protože při jejich použití testy nedodržují předepsanou hladinu.

Klíčová slova: Testování normality, lineární model, srovnání síly testů

Abstract

Title: Testing for Normality in Linear Model

Author: Blanka Hamplová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Karel Zvára, CSc.

Supervisor's e-mail address: karel.zvara@mff.cuni.cz

Abstract: This diploma thesis is concerned with the means of verifying the assumption that the random component of the dependent variable in a linear model is normally distributed. Within the realm of a simulation study we observed how the use of various types of residuals in place of the original vector of errors impacts the ability of ten tests of normality to adhere to their significance levels. Because the individual residuals depend on the matrix of the model, four specific matrices of frequently used models were considered. In addition, the impact of the choice of a model along with the type of residuals on the power against five different alternative distributions of individual tests was studied. It turns out that the behavior of the tests is governed more by the type of the residuals used than by the choice of the model. In particular, studentized residuals appear to be unsuitable, because with their use the tests do not adhere to the prescribed significance level.

Keywords: Testing for Normality, Linear Model, Comparison of Power

Úvod

V matematické statistice se často snažíme vysvětlit závislost naměřené vektorové veličiny na několika předem daných nezávisle proměnných (regresorech). K modelování této závislosti používáme lineární model. O naměřeném vektoru předpokládáme, že je tvořen nezávislými normálně rozdělenými náhodnými veličinami se stejným rozptylem a střední hodnotou závislou na regresní matici. Předpoklad normality je podstatný pro platnost testových statistik pro hypotézy týkající se například významnosti regresorů či platnosti podmodelu.

V praxi při ověřování předpokladu normality musíme aplikovat testy na rezidua místo na skutečný vektor chyb. V této práci se tedy budeme zabývat vlivem použití různých typů reziduí a jednotlivých běžně používaných modelů na chování zkoumaných testů normality. Do studie jsou zahrnuty i dva méně známé testy normality (Michaelův test D_{SP} a test T_3 od Ghosh), které jsou původně navrženy jako grafické testy.

Pro simulační studii, byl použit volně šiřitelný statistický software R 2.6.1 a jeho knihovny (k dispozici na www.r-project.org). Zdrojové kódy, data použitá pro simulace a výsledky simulací jsou na přiloženém CD.

1 Lineární model

Nechť $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ je vektor, jehož složky jsou nezávislé náhodné veličiny, a $\mathbf{X}_{n \times (k+1)} = (x_{ij})$ je známá matice konstant (regresní matice) pro jejíž rozměry platí $n > k + 1$. Pokud $h(\mathbf{X}) = r = k + 1$ mluvíme o modelu s plnou hodnotí. V případě $h(\mathbf{X}) = r < k + 1$ potom o modelu s neúplnou hodnotí. V této práci budeme nadále uvažovat pouze modely s plnou hodnotí.

Předpokládáme, že i -tou náhodnou veličinou Y_i lze zapsat ve tvaru:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + e_i, \quad i = 1, \dots, n.$$

Parametry $\beta_0, \beta_1, \dots, \beta_k$ jsou neznámé regresní koeficienty, x_{i1}, \dots, x_{ik} jsou prvky regresní matice tvořící i -tý řádek a e_1, \dots, e_n jsou nezávislé náhodné veličiny s následujícími vlastnostmi:

$$\begin{aligned} \mathbb{E} e_i &= 0, \\ \text{var } e_i &= \sigma^2, \quad i = 1, \dots, n. \end{aligned}$$

Rozptyl $\sigma^2 > 0$ je neznámý parametr. Tento lineární model můžeme zapsat také maticově ve tvaru

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}. \quad (1)$$

Vektor $\mathbf{X}\beta$ je nenáhodný a protože platí

$$\begin{aligned} \mathbb{E} \mathbf{e} &= \mathbf{0}, \\ \text{var } \mathbf{e} &= \sigma^2 \mathbf{I}_{n \times n}, \end{aligned}$$

má vektor \mathbf{Y} charakteristiky:

$$\begin{aligned} \mathbb{E} \mathbf{Y} &= \mathbf{X}\beta, \\ \text{var } \mathbf{Y} &= \sigma^2 \mathbf{I}_{n \times n}. \end{aligned}$$

V regresi se často snažíme odhadnout střední hodnotu $\mathbb{E}\mathbf{Y} = \mathbf{X}\beta$. Ta náleží do prostoru lineárních kombinací sloupců matice \mathbf{X} . Označme si tento prostor $\mathcal{M}(\mathbf{X})$. Nechť sloupce matice \mathbf{Q} tvoří nějakou ortonormální bázi prostoru $\mathcal{M}(\mathbf{X})$ a nechť sloupce matice \mathbf{N} doplní tuto bázi na ortonormální bázi prostoru \mathbf{R}^n . Dále si označme $\mathbf{H} = \mathbf{Q}\mathbf{Q}^T$ a $\mathbf{M} = \mathbf{N}\mathbf{N}^T$. Obě matice jsou symetrické, idempotentní a ve vzájemném vztahu

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

\mathbf{H} je projekční matice, protože obecnému vektoru \mathbf{y} přiřazuje jeho projekci $\mathbf{H}\mathbf{y}$ do prostoru $\mathcal{M}(\mathbf{X})$. Matici \mathbf{H} můžeme vyjádřit i pomocí regresní matice \mathbf{X} předpisem

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

Označme si $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ projekci vektoru \mathbf{Y} do $\mathcal{M}(\mathbf{X})$. Pro $\hat{\mathbf{Y}}$ platí

$$\begin{aligned} \mathbb{E} \hat{\mathbf{Y}} &= \mathbf{X}\beta, \\ \text{var} \hat{\mathbf{Y}} &= \sigma^2 \mathbf{H} \end{aligned}$$

a lze ukázat, že se jedná o nejlepší nestranný lineární odhad střední hodnoty $\mathbf{X}\beta$ (viz [14]).

Dále si označme

$$\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Tento vektor nazýváme vektorem reziduí. Vyjadřuje, nakolik se od sebe liší naměřené hodnoty vysvětlované veličiny a odhad jejich střední hodnoty na základě modelu. Můžeme psát

$$\mathbf{u} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

a díky (2) dostáváme

$$\mathbf{u} = \mathbf{M}\mathbf{Y}. \quad (3)$$

Matice \mathbf{M} je symetrická (tedy $\mathbf{M}^T = \mathbf{M}$) a její sloupce jsou kolmé k prostoru $\mathcal{M}(\mathbf{X})$, z čehož také plyne $\mathbf{M}\mathbf{X} = \mathbf{0}$. Tudíž lze vyjádření vektoru reziduí (3) přepsat do tvaru

$$\mathbf{u} = \mathbf{M}\mathbf{Y} = \mathbf{M}(\mathbf{X}\beta + \mathbf{e}) = \mathbf{M}\mathbf{e}. \quad (4)$$

Vektor reziduí má vlastnosti:

$$\begin{aligned} \mathbb{E} \mathbf{u} &= \mathbf{0}, \\ \text{var} \mathbf{u} &= \sigma^2 \mathbf{M}. \end{aligned}$$

Veličina

$$RSS = \|\mathbf{u}\|^2 = \sum_{i=1}^n u_i^2$$

se nazývá reziduální součet čtverců. Z něj je odvozen i nestranný odhad rozptylu v modelu (1)

$$S^2 = \frac{RSS}{n - r}, \quad (5)$$

který také bývá nazýván reziduální rozptyl.

Z vyjádření (4) je patrné, že vektor reziduí nezávisí na vektoru regresních koeficientů β . Díky tomu nám v simulacích ke spočítání vektoru reziduí stačí znát pouze vektor \mathbf{e} a matici modelu \mathbf{X} .

Všechny výše zmíněné vlastnosti platí v lineárním modelu bez ohledu na typ rozdělení vektoru \mathbf{Y} . Pro další úlohy (např. testování významnosti regresorů) musíme přidat ještě předpoklad normálního rozdělení vektoru \mathbf{Y} .

1.1 Normální lineární model

Jak už bylo uvedeno výše, uvažujeme pouze model s plnou hodnotí. V normálním lineárním modelu předpokládáme, že vektor pozorování \mathbf{Y} má mnohorozměrné normální rozdělení

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}). \quad (6)$$

To znamená, že chybový vektor \mathbf{e} má rozdělení $N(\mathbf{0}, \sigma^2\mathbf{I})$. Z vyjádření (3) je patrné, že vektor reziduí \mathbf{u} je lineární funkcí vektoru \mathbf{e} . Tudíž platí, že vektor reziduí \mathbf{u} má mnohorozměrné normální rozdělení $N(\mathbf{0}, \sigma^2\mathbf{M})$. Dále lze ukázat (viz [14]), že v normálním lineárním modelu má veličina $\frac{RSS}{\sigma^2}$ rozdělení χ_{n-r}^2 .

1.2 Další rezidua

Pro definici dalších reziduí potřebujeme předpoklad $m_{tt} > 0, t = 1, \dots, n$, tzn. předpoklad, že všechny diagonální prvky matice \mathbf{M} jsou kladné. V následujícím textu tedy budeme uvažovat pouze modely s výše zmíněnou vlastností. Podle [14] navíc platí

$$m_{tt} > 0 \Leftrightarrow h(\mathbf{X}_{[-t\bullet]}) = h(\mathbf{X}),$$

kde $\mathbf{X}_{[-t\bullet]}$ označuje matici \mathbf{X} s vynechaným t -tým řádkem.

1.2.1 Normovaná rezidua

Normovaná rezidua definujeme předpisem

$$v_i = \frac{u_i}{S\sqrt{m_{ii}}}, \quad i = 1, \dots, n, \quad (7)$$

kde S^2 je odhad rozptylu (5). Tato rezidua mají v normálním lineárním modelu vlastnosti (viz [14]):

$$\begin{aligned} \mathbb{E} v_i &= 0, \\ \text{var } v_i &= 1, \quad i = 1, \dots, n. \end{aligned}$$

V prostředí R můžeme normovaná rezidua spočítat příkazem `rstandard(a)`, kde `a` je výsledek použití funkce `lm()`.

1.2.2 Studentizovaná rezidua

Označme si

$$\mathbf{Y}_{[-t\bullet]} = \mathbf{X}_{[-t\bullet]}\beta + \mathbf{e}_{[-t\bullet]}$$

model (1) s vynechaným t -tým řádkem (pozorováním) a $\mathbf{b}_{[-t\bullet]}$ řešení normální rovnice v tomto modelu. Dále si označme $\mathbf{x}_{t\bullet}^T$ t -tý řádek matice \mathbf{X} . Odhady \mathbf{b} a $\mathbf{b}_{[-t\bullet]}$ jsou potom (viz [14]) ve vztahu

$$\mathbf{b}_{[-t\bullet]} = \mathbf{b} - \frac{u_t}{m_{tt}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{t\bullet}.$$

Odhad střední hodnoty t -tého pozorování spočtený na základě všech ostatních pozorování se dá zapsat ve tvaru $\mathbf{x}_{t\bullet}^T\mathbf{b}_{[-t\bullet]}$. Porovnání tohoto odhadu se skutečnou hodnotou t -tého pozorování můžeme zapsat

$$Y_t - \mathbf{x}_{t\bullet}^T\mathbf{b}_{[-t\bullet]} = Y_t - \mathbf{x}_{t\bullet}^T\left(\mathbf{b} - \frac{u_t}{m_{tt}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{t\bullet}\right) = \frac{u_t}{m_{tt}}. \quad (8)$$

Statistika (8) má nulovou střední hodnotu a rozptyl

$$\text{var}\left(\frac{u_t}{m_{tt}}\right) = \frac{\sigma^2 m_{tt}}{m_{tt}^2} = \frac{\sigma^2}{m_{tt}}.$$

Pokud při studentizaci statistiky (8) použijeme místo rozptylu σ^2 jeho nestranný odhad $S_{[-t\bullet]}^2$ z modelu s vynechaným t -tým pozorováním, dostaneme studentizované reziduum

$$v_t^* = \frac{u_t}{S_{[-t\bullet]}\sqrt{m_{tt}}}.$$

V normálním lineárním modelu (6) má studentizované reziduum v_t^* Studentovo t -rozdělení s $n - r - 1$ stupni volnosti (viz [14]).

Pro odhad rozptylu v modelu s vynechaným pozorováním platí

$$S_{[-t\bullet]}^2 = \frac{RSS_{[-t\bullet]}}{n - r - 1} = \frac{RSS - \frac{u_t^2}{m_{tt}}}{n - r - 1}.$$

Z porovnání $S_{[-t\bullet]}^2$ s odhadem rozptylu z původního modelu

$$\frac{S_{[-t\bullet]}^2}{S^2} = \frac{RSS - \frac{u_t^2}{m_{tt}}}{n - r - 1} \cdot \frac{n - r}{RSS} = \frac{n - r - v_t^2}{n - r - 1} \quad (9)$$

potom vyplývá

$$S_{[-t\bullet]}^2 < S^2 \Leftrightarrow |v_t| > 1.$$

Na základě (9) platí mezi normovaným reziduem (7) a studentizovaným reziduem (8) následující vztah:

$$v_t^* = v_t \left(\frac{n - r - 1}{n - r - v_t^2} \right)^{\frac{1}{2}},$$

který je užitečný pro výpočty.

V prostředí R získáme studentizovaná rezidua příkazem `rstudent(a)`, kde `a` je výsledek použití funkce `lm()`.

1.2.3 Nekorelovaná rezidua

Obě výše zmíněné transformace původních reziduí \mathbf{u} odstraňují problém nestejných rozptylů. Na rozdíl od vektoru chyb \mathbf{e} jsou však tato rezidua stále závislá.

Protože vektor reziduí leží v prostoru $\mathcal{M}(X)^\perp$ můžeme ho zapsat pomocí matice \mathbf{N} , jejíž sloupce tvoří ortonormální bázi prostoru $\mathcal{M}(X)^\perp$ (tj. pro kterou platí $\mathbf{N}^T \mathbf{N} = \mathbf{I}$ a $\mathbf{N} \mathbf{N}^T = \mathbf{M}$), ve tvaru

$$\mathbf{u} = \mathbf{N}(\mathbf{N}^T \mathbf{Y}) = \mathbf{N} \mathbf{n}.$$

Vektor \mathbf{n} je potom vektor nekorelovaných reziduí. V normálním modelu (6) jsou tato rezidua nezávislá a platí pro ně (viz [14])

$$\begin{aligned} \mathbf{E} \mathbf{n} &= \mathbf{E}(\mathbf{N}^T \mathbf{Y}) = \mathbf{N}^T \mathbf{X} \beta = \mathbf{0}, \\ \text{var} \mathbf{n} &= \text{var}(\mathbf{N}^T \mathbf{Y}) = \text{var}(\mathbf{N}^T \mathbf{e}) = \sigma^2 \mathbf{I}_{n-r}. \end{aligned}$$

Protože existuje nekonečně mnoho ortonormálních bází prostoru $\mathcal{M}(X)^\perp$, není matice \mathbf{N} dána jednoznačně. My se zaměříme na matici \mathbf{N} , která určuje tzv. rekurzivní rezidua. Při její konstrukci budeme postupovat následovně: Začneme prvním řádkem matice \mathbf{X} a postupně budeme přidávat další. V každém kroku, kdy se nezvýší hodnota rozšiřované matice, spočteme rozdíl mezi přidávaným pozorováním a odhadem jeho střední hodnoty založeným na všech doposud přidávaných pozorováních. Tento rozdíl ještě znormujeme tak, aby vzniklá statistika měla rozptyl σ^2 . Označme si $(\mathbf{Y}_t, \mathbf{X}_t)$ matici tvořenou prvními t řádky matice (\mathbf{Y}, \mathbf{X}) , hodnota příslušné regresní matice $h(\mathbf{X}_t) = r_t$ a \mathbf{b}_t řešení normální rovnice příslušné k \mathbf{X}_t . Předpokládejme, že po přidání dalšího řádku $(Y_{t+1}, (\mathbf{x}_{t+1, \bullet})^T)$ platí $h(\mathbf{X}_t) = h(\mathbf{X}_{t+1}) = r_t$, tj. že se nezvýší hodnota. Potom můžeme psát

$$n_{t-r_t+1} = \frac{Y_{t+1} - (\mathbf{x}_{t+1, \bullet})^T \mathbf{b}_t}{\sqrt{1 + (\mathbf{x}_{t+1, \bullet})^T (\mathbf{X}_t^T \mathbf{X}_t)^{-1} \mathbf{x}_{t+1, \bullet}}}. \quad (10)$$

Postupem (10) dostaneme postupně statistiky n_1, \dots, n_{n-r} , z nichž každá je nekorelovaná se všemi statistikami s nižším indexem (viz [14]). Pokud má smysl uspořádání řádků matice (\mathbf{Y}, \mathbf{X}) , můžeme nekorelovaná rezidua interpretovat jako srovnání, nakolik každé další pozorování odpovídá modelu zkonstruovanému na základě všech předchozích pozorování.

2 Testy normality

V této sekci uvádím přehled použitých testů normality. Všechny testy byly vytvořeny pro rozhodnutí, zda daný náhodný výběr pochází z normálního rozdělení. V případě lineárního modelu můžeme za náhodný výběr považovat pouze chybový vektor \mathbf{e} . Ten ovšem nikdy nepozorujeme, proto testy normality v praxi používáme na vektory reziduí, která už však nemají vlastnosti náhodného výběru. Je proto zajímavé sledovat vliv použití reziduí na dodržování hladiny a sílu testů.

Obecně známým diagnostickým prostředkem používaným pro detekci odchylek od normality je normální diagram, který je založen na následující myšlence:

Nechť G_1, \dots, G_n je výběr z rozdělení $N(0, 1)$. Označme si

$$g_i = E G_{(i)}, \quad i = 1, \dots, n.$$

Pro uspořádaný výběr $X_{(1)}, \dots, X_{(n)}$ z rozdělení $N(\mu, \sigma^2)$ potom platí

$$E X_{(i)} = \mu + \sigma g_i, \quad i = 1, \dots, n. \quad (11)$$

Body $[g_i, X_{(i)}]$ by tudíž měly být rozloženy kolem přímky $y = \mu + \sigma x$. Diagram normality je grafický prostředek, který nám umožňuje pouze subjektivní hodnocení, zda náš výběr pochází z normálního rozdělení či nikoli. Pokud se tedy tvar diagramu normality zkoumaného výběru výrazněji liší od přímky (je konvexní, konkávní, či esovitě prohnutý), stačí nám k zamítnutí hypotézy normality pouhý pohled. V případě, že si nejsme jisti, je vhodné použít některý z objektivních statistických testů.

2.1 Shapirův-Wilkův test W

Tento test zveřejněný v [10] úzce souvisí s výše zmíněným diagramem normality. Shapiro a Wilk jej založili na porovnání dvou odhadů parametru σ . Označme si

$$s_*^2 = (n-1)s^2 = \sum_{i=1}^n (X_i - \bar{X})^2$$

statistiku vycházející z odhadu parametru σ pomocí výběrového rozptylu pozorování $X_{(1)}, \dots, X_{(n)}$.

Pokud je $X_{(1)}, \dots, X_{(n)}$ uspořádaný náhodný výběr z rozdělení $N(0,1)$, můžeme druhý odhad parametru σ spočítat z modelu

$$E \mathbf{X}_0 = \mu \mathbf{1} + \sigma \mathbf{g}, \quad (12)$$

který je pouze maticovým zápisem vztahu (11). Zde je $\mathbf{X}_() = (X_{(1)}, \dots, X_{(n)})^T$ a $\mathbf{g} = (g_1, \dots, g_n)^T$ označuje vektor středních hodnot uspořádaného náhodného výběru z normovaného normálního rozdělení. Dále si označíme \mathbf{V} varianční matici spořádaného náhodného výběru z rozdělení $N(0,1)$. Odhad parametru σ (tedy směrnice přímký v diagramu normality) získaný zobecněnou metodou nejmenších čtverců z modelu (12) je potom

$$\hat{\sigma} = \frac{\mathbf{g}^T \mathbf{V}^{-1} \mathbf{X}_()}{\mathbf{g}^T \mathbf{V}^{-1} \mathbf{g}}.$$

Zavedeme si koeficienty $\mathbf{a} = (a_1, \dots, a_n)^T$ dané předpisem

$$\mathbf{a} = \frac{\mathbf{g}^T \mathbf{V}^{-1}}{\|\mathbf{g}^T \mathbf{V}^{-1}\|},$$

které mají vlastnosti $a_i = -a_{n-i+1}$ a $\mathbf{a}^T \mathbf{a} = 1$. Tudíž můžeme statistiku Shapirova–Wilkova testu

$$W = \frac{\hat{\sigma}^2 (\mathbf{g}^T \mathbf{V}^{-1} \mathbf{g})^2}{\|\mathbf{g}^T \mathbf{V}^{-1}\|^2 s_*^2} = \frac{(\mathbf{a}^T \mathbf{X}_())^2}{s_*^2}$$

zapsat ve tvaru:

$$W = \frac{\left[\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} a_i (X_{(i)} - X_{(n-i+1)}) \right]^2}{s_*^2}.$$

Protože výpočet koeficientů \mathbf{a} vyžaduje znalost varianční matice \mathbf{V} uspořádaného náhodného výběru z $N(0,1)$, navrhli autoři v [10] nahradit koeficienty \mathbf{a} jejich aproximací $\hat{\mathbf{a}}^*$. Výpočet $\hat{\mathbf{a}}^*$ je založen pouze na znalosti vektoru středních hodnot uspořádaného náhodného výběru z $N(0,1)$ a rozsahu výběru n . Původně byly kritické hodnoty pro W tabelovány pouze pro rozsahy výběru $n < 50$. Tento nedostatek odstranil Royston v článku [6], kde navrhl normalizující transformaci statistiky W , která umožňovala použití Shapirova–Wilkova testu pro rozsahy výběru $7 \leq n \leq 2000$ a výpočet p-hodnot.

V R je Shapirův–Wilkův test k dispozici pod názvem `shapiro.test()`. Pro výpočet využívá Roystonův upravený algoritmus zveřejněný v [8], který je možné použít pro rozsahy výběru $3 \leq n \leq 5000$.

2.2 Shapirův–Franciův test W'

Shapiro a Francia v [9] navrhli zjednodušení statistiky W . Nahradili koeficienty \mathbf{a} jejich modifikací

$$\mathbf{a}^* = \frac{\mathbf{g}}{\sqrt{\mathbf{g}^T \mathbf{g}}},$$

kteřá nevyžaduje výpočet varianční matice \mathbf{V} uspořádaného náhodného výběru z normovaného normálního rozdělení. Testová statistika má potom tvar

$$W' = \frac{[\mathbf{g}^T \mathbf{X}_0]^2}{\|\mathbf{g}\|^2 \|\mathbf{X}_0 - \bar{X}\|^2} = \frac{[\sum_{i=1}^n g_i X_{(i)}]^2}{\sum_{i=1}^n g_i^2 \sum_{i=1}^n (X_{(i)} - \bar{X})^2}.$$

Jedná se tedy vlastně o čtverec výběrového korelačního koeficientu $r_{\mathbf{g}, \mathbf{X}_0}^2$. Protože původně byly kritické hodnoty Shapirova-Wilkova testu W tabelovány pouze pro rozsahy výběru $n < 50$, doporučovali Shapiro a Francia pro větší rozsahy používat jako aproximaci test W' .

V prostředí R je pro vektor středních hodnot uspořádaného náhodného výběru z rozdělení $N(0, 1)$ použita aproximace

$$g_i = \Phi^{-1} \left(\frac{i - a}{n + (1 - a) - a} \right),$$

kde

$$a = \begin{cases} \frac{3}{8}, & n \leq 10 \\ \frac{1}{2}, & n > 10. \end{cases}$$

Tento test nachází v knihovně `nortest` jako funkce `sf.test()` a lze ho použít pro rozsahy výběru $5 \leq n \leq 5000$.

2.3 Test založený na šikmosti ($\sqrt{b_1}$)

Označme si i -tý centrální moment pravděpodobnostního rozdělení

$$\mu_i = E(X - EX)^i.$$

Šikmost $\sqrt{\beta_1}$ je potom dána předpisem

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}.$$

Definujeme si výběrovou šikmost

$$\sqrt{b_1} = \frac{m_3}{m_2^{\frac{3}{2}}},$$

kde

$$m_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k.$$

Pro normální rozdělení platí

$$E \sqrt{b_1} = 0, \quad \text{var} \sqrt{b_1} = \frac{6(n-2)}{(n+1)(n+3)}.$$

Dříve se používal test založený na statistice

$$Y = \sqrt{b_1} \left(\frac{(n+1)(n+3)}{6(n-2)} \right)^{\frac{1}{2}},$$

která má asymptoticky normální rozdělení $N(0, 1)$. Tato asymptotika je však pomalá a tudíž nevhodná pro malé rozsahy výběrů. Autoři článku [3] doporučují použít následující postup pro transformaci veličiny Y :

Spočteme statistiky

$$\beta_2(\sqrt{b_1}) = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

$$W^2 = -1 + \left(2(\beta_2(\sqrt{b_1}) - 1) \right)^{\frac{1}{2}}, \quad \delta = \frac{1}{\sqrt{\ln W}},$$

$$\alpha = \left(\frac{2}{W^2 - 1} \right)^{\frac{1}{2}}.$$

Transformovaná veličina

$$Z(\sqrt{b_1}) = \delta \ln \left[\frac{Y}{\alpha} + \left(\left(\frac{Y}{\alpha} \right)^2 + 1 \right)^{\frac{1}{2}} \right]$$

má potom za platnosti nulové hypotézy přibližně normální rozdělení $N(0,1)$. Tento test lze použít pro rozsahy výběrů $n > 8$.

2.4 Test založený na špičatosti (b_2)

Při zachování značení z předchozího odstavce si můžeme definovat špičatost rozdělení β_2 předpisem

$$\beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Dále si definujeme výběrovou špičatost

$$b_2 = \frac{m_4}{m_2^2}.$$

Za platnosti hypotézy normality je její střední hodnota

$$E(b_2) = \frac{3(n-1)}{n+1}$$

a rozptyl

$$\text{var}(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Standardizovanou verzi statistiky b_2 si označíme

$$Y = \frac{b_2 - E(b_2)}{\text{var}(b_2)}.$$

Statistika Y má sice asymptoticky normální rozdělení $N(0, 1)$, ale tato asymptotika je ještě pomalejší než v případě výběrové šikmosti. Autoři článku [3] doporučují následující postup pro normalizaci Y :

Třetí standardizovaný moment b_2 je

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2 - 5n + 2)}{(n + 7)(n + 9)} \sqrt{\frac{6(n + 3)(n + 5)}{n(n - 2)(n - 3)}}.$$

Definujeme si veličinu A

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}} \left[\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{\left(1 + \frac{4}{\beta_1(b_2)}\right)} \right],$$

pomocí které provedeme transformaci výběrové špičatosti b_2

$$Z(b_2) = \frac{\left(\left(1 - \frac{2}{9A}\right) - \left[\frac{1 - \frac{2}{A}}{1 + Y \sqrt{\frac{2}{A-4}}} \right]^{\frac{1}{3}} \right)}{\sqrt{\frac{2}{9A}}}.$$

$Z(b_2)$ má za platnosti nulové hypotézy přibližně normální rozdělení $N(0, 1)$. Tento test je doporučen pro rozsahy výběru $n \geq 20$.

2.5 Souhrnný test K^2

Souhrnný test založený na výběrové šikmosti i špičatosti je kombinací předchozích dvou testů. Testová statistika

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)$$

má za platnosti nulové hypotézy normality přibližně rozdělení χ_2^2 . Tento test lze použít pro rozsahy $n \geq 20$.

2.6 Test T_3 založený na EMVF

Test T_3 vytvořila Sucharita Ghosh [4] a je založen na třetí derivaci logaritmu empirické momentové vytvořující funkce. Momentová vytvořující funkce je definována předpisem

$$\mu(t) = E(e^{tX}).$$

Mějme náhodný výběr X_1, \dots, X_n . Empirická momentová vytvořující funkce je potom dána předpisem

$$M(t) = \frac{1}{n} \sum_{j=1}^n \exp\{tX_j\}.$$

Definujeme si empirickou momentovou vytvořující funkci pro studentizovaná data

$$\tilde{M}(t) = \frac{1}{n} \sum_{j=1}^n \exp\left\{\frac{t(X_j - \bar{X})}{s}\right\}$$

a její derivaci podle t

$$\begin{aligned} \tilde{M}^{(i)}(t) &= \frac{d^i}{dt^i} \tilde{M}(t) = \frac{1}{n} \sum_{j=1}^n \left(\frac{(X_j - \bar{X})}{s}\right)^i \exp\left\{\frac{t(X_j - \bar{X})}{s}\right\}, \quad i > 0 \\ &= 0, \quad i < 0 \\ \tilde{M}^{(0)}(t) &= M(t). \end{aligned}$$

Pro rozsah výběru n si označíme třetí derivaci logaritmu empirické momentové vytvořující funkce

$$\begin{aligned} T_3^{(n)}(t) &= \sqrt{n} \frac{d^3}{dt^3} \{\ln M(t)\} \\ &= \sqrt{n} M^{-2}(t) \left\{ M^{(3)}(t) M(t) - 3M^{(2)}(t) M^{(1)}(t) + 2 \frac{M^{(2)}(t)}{M(t)} \right\}. \end{aligned}$$

Nechť $Z(t)$ je gaussianský proces s nulovou střední hodnotou a kovarianční funkcí $\text{cov}\{Z(t), Z(s)\} = \gamma(t, s)$ definovaný pro $t \in [-a, a]$. Nechť m je medián $\sup_{|t| \leq a} |Z(t)|$ a $\delta = \sup_{|t| \leq a} \{\gamma(t, t)\}$. Potom platí

$$P\left\{\sup_{|t| \leq a} |Z(t)| \leq m + u_\alpha \delta\right\} \geq 1 - \alpha. \quad (13)$$

Podle [4] může být tato nerovnost použita na standardizovanou statistiku

$$\tilde{T}_3^{(n)}(t) = \frac{T_3^{(n)}(t)}{\sqrt{K_3(t, t)}},$$

kde $K_3(t, t)$ je asymptotický rozptyl odvozený v [4]. V našem případě je $\delta = 1$ a m můžeme aproximovat mediánem $m_3(n)$ získaným ze simulací. Jeho vyjádření jako funkce rozsahu výběru n je ve zmíněném článku uvedeno takto:

$$\hat{m}_3(n) = 1,38 - 5,92 \frac{1}{\sqrt{n}} + 12,8 \frac{1}{n}. \quad (14)$$

V testu, který autorka naprogramovala pro prostředí S-plus, a který mi laskavě poskytla, je na základě dalších simulací odhad mediánu lehce pozměněn do tvaru:

$$\hat{m}_3(n) = 1,217338 - 3,568419 \frac{1}{\sqrt{n}} + 4,59114 \frac{1}{n}. \quad (15)$$

Použitím vztahů (13) a (15) dostáváme přibližný $(1 - \alpha) \cdot 100\%$ kvantil pro statistiku $\tilde{T}_3^{(n)}(t)$

$$\tilde{q}(\alpha, n) = \hat{m}_3(n) + u_\alpha.$$

Tento test byl původně vytvořen jako grafický test. Nulovou hypotézu normality tedy zamítáme, pokud graf funkce $T_3^{(n)}(t)$ (křivka proložená body $T_3^{(n)}(t), t \in \{0, \pm 0.5, \pm 0.1, \dots, \pm 1\}$) kdekoli na intervalu $[-1, 1]$ protne křivky, které jsou dány předpisem $y = \pm \hat{m}_3(n) \sqrt{K_3(t, t)}$. Pokud k tomu dojde, můžeme z chování funkce $T_3^{(n)}(t)$ v okolí nuly odhadnout, k jakému porušení normality došlo. Pro normální rozdělení je totiž graf funkce $T_3^{(n)}(t)$ přímka procházející bodem $y = 0$. Posunutí ve směru osy y v bodě $t = 0$ naznačuje šikmost různou od 0. Nenulová směrnice grafu ve stejném bodě potom odpovídá špičatosti různé od 3.

T3 test byl také implementován v prostředí R a to v knihovně `cwhmisc` (v balíku `cwhplot`) jako funkce `T3plot(x)`. Autor této knihovny však používá pro výpočet kritických hodnot odhad mediánu

$$\hat{m} = 0,7168311 - 2,327602 \frac{1}{\sqrt{n}} + 3,688362 \frac{1}{n}, \quad (16)$$

který je odlišný od (15) i (14). Tato „kosmetická úprava“ (jak ji autor sám nazývá) odhadu mediánu vede k přesnějšímu dodržování hladiny testu. Proto jsem v této práci použila (16).

2.7 Lillieforsův test D^*

V knize [13] je zmíněna Lillieforsova modifikace Kolmogorovova-Smirnovova testu, který je založen na porovnání empirické distribuční funkce náhodného výběru s distribuční funkcí rozdělení za nulové hypotézy ($F_0(x)$). Empirická distribuční funkce náhodného výběru X_1, \dots, X_n je

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}_{[X_i \leq x]}.$$

Statistika Kolmogorovova-Smirnovova testu má tvar

$$D = \max \{ D^+, D^- \},$$

kde

$$D^+ = \max_{i=1, \dots, n} \left\{ \frac{i}{n} - F_0(X_{(i)}) \right\},$$

$$D^- = \max_{i=1, \dots, n} \left\{ F_0(X_{(i)}) - \frac{i-1}{n} \right\}.$$

Jedná se tedy o maximum rozdílů empirické distribuční funkce a distribuční funkce z nulové hypotézy. Kolmogorovův-Smirnovův test je testem jednoduché hypotézy, proto musí být specifikovány parametry rozdělení μ a σ . Lilliefors na základě simulací aproximoval rozdělení statistiky D pro složenou nulovou hypotézu, že výběr pochází z obecného normálního rozdělení $N(\mu, \sigma^2)$, a tabeloval kritické hodnoty pro různé rozsahy výběrů. Stephens v [11] navrhuje zjednodušující transformaci

$$D^* = D \left(\sqrt{n} - 0,01 + \frac{0,85}{\sqrt{n}} \right),$$

která umožňuje pro danou hladinu α použít stejnou kritickou hodnotu pro všechny rozsahy výběru. Takto modifikovaný Lillieforsův test se v prostředí R nachází v knihovně `nortest` jako funkce `lillie.test()` a je možné ho použít pro rozsahy výběru $n \geq 5$.

2.8 Michaelův test D_{SP}

John R. Michael v článku [5] prezentoval grafický test vycházející z Kolmogorova-Smirnova testu, který rozšiřuje možnosti normálního diagramu.

Mějme $X_{(1)}, \dots, X_{(n)}$ uspořádaný náhodný výběr. Definujeme si body

$$f_i = \Phi \left(\frac{X_{(i)} - \bar{X}}{\tilde{s}} \right), \quad (17)$$

kde $\Phi(z)$ označuje distribuční funkci normovaného normálního rozdělení, \bar{X} je výběrový průměr a

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Dále si označíme transformaci stabilizující rozptyl

$$g(x) = \frac{2}{\pi} \arcsin(\sqrt{x}).$$

Michaelova statistika D_{SP} je potom dána předpisem

$$D_{SP} = \max |g(f_i) - g(p_i)|,$$

kde

$$p_i = \frac{i - \frac{1}{2}}{n}.$$

V článku [7] Royston popisuje normalizační transformaci Michaelovy statistiky D_{SP} . Pomocí simulací byl aproximován medián \tilde{D} rozdělení D_{SP} za platnosti nulové hypotézy jako funkce rozsahu výběru n :

$$\tilde{D} = 0,016264n^{-\frac{1}{4}} + 0,59296n^{-\frac{1}{2}} - 0,58067n^{-\frac{3}{4}}$$

a odvozeny koeficienty

$$\begin{aligned}\tilde{\lambda} &= 0,34963 \ln(n) - 0,10246(\ln(n))^2 + 0,00095390(\ln(n))^4, \\ \tilde{\beta} &= 0,22153 + 0,46995n^{-0,86}\end{aligned}$$

tak, že transformovaná statistika

$$z = \frac{\left(\frac{D_{SP}}{\tilde{D}}\right)^{\tilde{\lambda}} - 1}{\tilde{\beta}\tilde{\lambda}} \quad (18)$$

má potom přibližně normované normální rozdělení. To nám dává návod pro výpočet kritických hodnot testu.

Michaelův test umožňuje do modifikovaného normálního diagramu zakreslit kritický obor testu. Nulovou hypotézu zamítáme, pokud některý z bodů

$$t_i = \frac{2}{\pi} \arcsin \left(\sqrt{\Phi \left(\frac{X_i - \bar{X}}{\tilde{s}} \right)} \right)$$

padne mimo pás určený předpisem

$$y = \frac{2}{\pi} \arcsin \left(\sqrt{\frac{i - \frac{1}{2}}{n}} \right) \pm D_\alpha.$$

D_α zde označuje $(1 - \alpha)\%$ kvantil spočtený pomocí (18). Tento test je doporučován pro rozsah výběru $7 \leq n \leq 1000$. Test D_{SP} je pro prostředí R naprogramován na příloženém CD jako funkce `D_SP.test()`.

2.9 Cramerův-von Misesův test W^{2*}

Cramerův-von Misesův test je uveden např. v [13]. Je jedním z rodiny kvadratických testů založených na porovnání empirické distribuční funkce a distribuční funkce z nulové hypotézy. Tyto testy mají společný tvar

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \psi(F(x)) dF(x), \quad (19)$$

kde $\psi(F(x))$ je váhová funkce.

Volbou váhové funkce $\psi(F(x)) = 1$ dostáváme Cramerovu-von Misesovu testovou statistiku (jejíž označení pouze připomíná Shapirovu-Wilkovu statistiku W):

$$W^2 = \frac{1}{12n} + \sum_{i=1}^n \left(p_{(i)} - \frac{2i-1}{2n} \right)^2,$$

kde

$$p_{(i)} = \Phi \left(\frac{X_{(i)} - \bar{X}}{s} \right). \quad (20)$$

Stephens ve svém článku [11] uvádí modifikaci

$$W^{2*} = \left(1 + \frac{1}{2n} \right) W^2,$$

která umožňuje použít pro požadovanou hladinu α stejnou kritickou hodnotu pro libovolný rozsah výběru. Takto upravený test je v prostředí R implementován v knihovně `nortest` jako funkce `cvm.test()` a je možné ho použít pro rozsah výběru $n \geq 8$.

2.10 Andersonův-Darlingův test A^{2*}

Andersonův-Darlingův test je další z rodiny kvadratických testů (19) zmíněných v [13] založených na empirické distribuční funkci. Použití váhové funkce

$$\psi(p) = \frac{1}{p(1-p)}$$

vede k testové statistice

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) (\ln(p_{(i)}) + \ln(1-p_{(n-i+1)})),$$

kde $p_{(i)}$ je definováno v (20).

V prostředí R je tento test implementován s modifikovanou testovou statistikou

$$A^{2*} = \left(1 + \frac{0,75}{n} + \frac{2,25}{n^2} \right) A^2,$$

která umožňuje použít pro požadovanou hladinu α stejnou kritickou hodnotu pro libovolný rozsah výběru. Můžeme ho nalézt v knihovně `nortest` jako funkci `ad.test()` a je možné ho použít pro rozsah výběru $n \geq 8$.

3 Rozdělení

Na tomto místě uvádím přehled pravděpodobnostních rozdělení použitých v simulační studii. Pokud je to možné, jsou parametry jednotlivých rozdělení voleny tak, aby měla střední hodnotu 0 a rozptyl 1. Pokud této vlastnosti nelze dosáhnout přímo volbou parametrů, použijeme jednoduchou lineární transformaci.

Protože je několik z použitých testů založeno na šikmosti nebo špičatosti, uvádím u jednotlivých rozdělení hodnoty těchto charakteristik. Pro jejich výpočet jsou užitečné tyto vzorce pro výpočet třetího a čtvrtého centrálního momentu (viz [1]):

$$\begin{aligned}\mu_3 &= E X^3 - 3E X^2 E X + 2(E X)^3, \\ \mu_4 &= E X^4 - 4E X^3 E X + 6E X^2 (E X)^2 - 3(E X)^3.\end{aligned}\quad (21)$$

Z definice šikmosti a špičatosti vyplývá, že na jejich hodnotu nemá lineární transformace původní náhodné veličiny žádný vliv. Proto i veličiny transformované tak, aby měly nulovou střední hodnotu a jednotkový rozptyl, budou mít stejné charakteristiky $\sqrt{\beta_1}$ a β_2 jako veličiny původní.

3.1 Normální rozdělení

Hladinu jednotlivých testů budeme ověřovat na výběru z normálního rozdělení $N(\mu, \sigma^2)$. To má hustotu

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbf{R}, \mu \in \mathbf{R}, \sigma^2 > 0.$$

Střední hodnota a rozptyl

$$E X = \mu, \quad \text{var } X = \sigma^2.$$

Šikmost a špičatost (viz [1])

$$\sqrt{\beta_1} = 0, \quad \beta_2 = 3.$$

Pro generování náhodného výběru z rozdělení $N(\mu, \sigma^2)$ o rozsahu n můžeme v prostředí R použít funkci `rnorm` (n, μ, σ). V simulacích používáme normální rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$.

3.2 Rovnoměrné rozdělení

Hustota spojitého rovnoměrného rozdělení $R(a, b)$ je:

$$f(x) = \frac{1}{b-a}, \quad a < x < b.$$

Střední hodnota a rozptyl

$$E X = \frac{a+b}{2}, \quad \text{var } X = \frac{(b-a)^2}{12}.$$

Obecný i -tý moment je potom

$$\mu'_i = \frac{b^{i+1} - a^{i+1}}{(i+1)(b-a)}.$$

Šikmost a špičatost

$$\sqrt{\beta_1} = 0, \quad \beta_2 = \frac{9}{5}.$$

Pro generování náhodného výběru z rozdělení $R(a, b)$ o rozsahu n můžeme v prostředí R použít funkci `runif(n, min=a, max=b)`. Rovnoměrné rozdělení s nulovou střední hodnotou a jednotkovým rozptylem dostaneme volbou parametrů $a = -\sqrt{3}$ a $b = \sqrt{3}$.

3.3 Dvojitě exponenciální rozdělení

Hustota dvojitě exponenciálního rozdělení $DEx(a, b)$ je:

$$f(x) = \frac{1}{2b} \exp\left\{-\frac{|x-a|}{b}\right\}, \quad x \in \mathbf{R}, a \in \mathbf{R}, b > 0.$$

Střední hodnota a rozptyl

$$E X = a, \quad \text{var } X = 2b^2.$$

Šikmost a špičatost

$$\sqrt{\beta_1} = 0, \quad \beta_2 = 6.$$

Pro toto rozdělení jsem v běžných knihovných prostředí R nenalezla funkci generující náhodný výběr. Ten lze získat následujícím postupem: Spočteme si distribuční funkci dvojitě exponenciálního rozdělení

$$F(x) = \int_{-\infty}^x f(z) dx = \frac{1}{2} \left[1 + \text{sign}(x - \mu) \left(1 - \exp\left\{-\frac{|x - \mu|}{b}\right\} \right) \right].$$

Dále k ní můžeme spočítat kvantilovou funkci

$$F^{-1}(z) = a - b \operatorname{sign} \left(z - \frac{1}{2} \right) \ln \left(1 - 2 \left| z - \frac{1}{2} \right| \right).$$

Budeme generovat náhodné veličiny U z rovnoměrného rozdělení $R(0, 1)$ (v prostředí R pomocí funkce `runif()`). Veličiny $F^{-1}(U)$ potom mají dvojitě exponenciální rozdělení. Abychom dostali náhodné veličiny s charakteristikami $E X = 0$ a $\operatorname{var} X = 1$, použijeme dvojitě exponenciální rozdělení s parametry $a = 0$ a $b = \frac{1}{\sqrt{2}}$.

3.4 Exponenciální rozdělení

Hustota rozdělení $Exp(\lambda)$ je:

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x > 0, \lambda > 0.$$

Střední hodnota a rozptyl

$$E X = \lambda, \quad \operatorname{var} X = \lambda^2.$$

Obecný i -tý moment je potom

$$\mu'_i = \lambda^i i!. \quad (22)$$

Za pomoci (21) můžeme z (22) spočítat šikmost a špičatost

$$\sqrt{\beta_1} = 2, \quad \beta_2 = 9.$$

Pro generování náhodného výběru z rozdělení $Exp(\lambda)$ o rozsahu n můžeme v prostředí R použít funkci `rexp(n, rate=1/lambda)`. Pro simulace použijeme rozdělení $Exp(2)$. Má-li náhodná veličina X rozdělení $Exp(2)$, potom transformovaná náhodná veličina $Y = \frac{X-2}{2}$ má nulovou střední hodnotu a jednotkový rozptyl.

3.5 Studentovo t -rozdělení

Hustota Studentova t -rozdělení t_n s n stupni volnosti je:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2}) \sqrt{\pi n}} \left(1 + \frac{x^2}{n} \right)^{-\frac{n+1}{2}}, \quad x > 0, n = 1, 2, \dots$$

Střední hodnota existuje pro $n > 1$ a je rovna nule. Pro $n > 2$ existuje konečný rozptyl

$$\text{var } X = \frac{n}{n-2}.$$

Šikmost a špičatost (viz [2])

$$\begin{aligned}\sqrt{(\beta_1)} &= 0, & n > 3 \\ \beta_2 &= \frac{3n-6}{n-4}, & n > 4\end{aligned}$$

Pro generování náhodného výběru z rozdělení t_n o rozsahu k můžeme v prostředí R použít funkci `rt(k,df=n)`. Pro simulace použijeme rozdělení t_5 . Jestliže má náhodná veličina X rozdělení t_5 , potom transformovaná náhodná veličina $Y = \sqrt{\frac{3}{5}}X$ má nulovou střední hodnotu a jednotkový rozptyl.

3.6 Smíšené rozdělení

Poslední použité rozdělení vychází z myšlenky dvou výběrů. Jedná se o dva výběry z normálního rozdělení, které se liší rozptylem i rozsahem. Pro požadovaný rozsah výběru n je tedy $\frac{2n}{5}$ pozorování náhodným výběrem z $N(0, 1)$ a $\frac{3n}{5}$ náhodným výběrem z $N(0, 4)$. Tyto rozsahy jsou voleny tak, aby korespondovaly s maticí modelu jednoduchého třídění použitou v simulační studii.

4 Simulace

Pro simulační studii jsme zvolili deset různých testů normality zmíněných v kapitole 2, čtyři matice modelu \mathbf{X} a pět různých modifikací chybového vektoru \mathbf{e} (\mathbf{e} , \mathbf{u} , \mathbf{v} , \mathbf{v}^* , \mathbf{n}). Vektor chyb \mathbf{e} jsme generovali z rozdělení zmíněných v kapitole 3. V celé simulační studii jsme pracovali s hladinou významnosti $\alpha = 0,05$.

Při volbě matic \mathbf{X} jsem se inspirovala diplomovou prací Mgr. Štefka [12], která byla zpracována na podobné téma. První matice odpovídá modelu regresní přímky. Je tvořena sloupcem jedniček (absolutní člen) a sloupcem náhodných čísel generovaných z rovnoměrného rozdělení se střední hodnotou 0 a s rozptylem 25, tj. z rozdělení $R(-8,66; 8,66)$. Má tedy 2 sloupce a 250 řádků. Druhá matice odpovídá modelu kvadratické závislosti. První sloupec matice tvoří samé jedničky, druhý opět náhodný výběr z rozdělení $R(-8,66; 8,66)$ a třetí sloupec je druhou mocninou druhého sloupce. Třetí matice je regresní matice odpovídající regresi s absolutním členem a třemi regresory. Má tedy rozměry 250×4 , kde první sloupec tvoří jedničky a zbylé tři sloupce jsou opět náhodně vygenerovaná čísla z rozdělení $R(-8,66; 8,66)$. Poslední matice odpovídá modelu nevyváženého jednoduchého třídění s rozsahy výběrů v poměru 2:3.

Pro každé z rozdělení jsme generovali 1000 vektorů \mathbf{e} délky 250. Simulace jsme prováděli pro rozsahy výběrů $n = 10, 15, 20, \dots, 250$. Jednotlivé rozsahy jsme dostali tak, že jsme vzali vždy prvních n složek vektoru \mathbf{e} . Analogicky k tomu jsme použili prvních n řádků matice \mathbf{X} . Pouze v případě modelu jednoduchého třídění jsme museli použít jiný postup při vytváření matice modelu. Tu jsme vytvářeli pro každý požadovaný rozsah výběru n zvlášť tak, aby první sloupec obsahoval na prvních $\frac{2n}{5}$ pozicích jedničky a zbylé pozice obsadily nuly. Druhý sloupec potom začíná $\frac{2n}{5}$ nulami, jež jsou doplněny jedničkami do celkové délky n .

Jak už bylo zmíněno v odstavci 1.2.3, nekorelovaná rezidua mají rozumnou interpretaci pouze pokud má uspořádání řádků matice modelu nějaký smysl. Protože tomu tak v případě jednoduchého třídění není, nebudeme v posledně zmiňovaném modelu nekorelovaná rezidua používat.

4.1 Ověření hladiny testů

Dodržování předepsané hladiny $\alpha = 5\%$ jednotlivými testy jsme ověřovali tak, že jsme je aplikovali na 1000 výběrů z normálního rozdělení, a sledovali počet zamítnutí hypotézy normality. Zamítnutí hypotézy můžeme popsat

veličinou

$$X_i = \begin{cases} 1, & \text{test zamítá hypotézu normality,} \\ 0, & \text{test nezamítá hypotézu normality,} \end{cases} \quad i = 1, \dots, 1000.$$

Předpokládáme, že veličiny $X_i, i = 1, \dots, 1000$ mají alternativní rozdělení s parametrem α a tudíž platí

$$P(X_i = 1) = \alpha, \quad P(X_i = 0) = 1 - \alpha.$$

Můžeme spočítat odhad parametru α

$$\hat{\alpha} = \frac{\sum_{i=1}^{1000} X_i}{1000}.$$

Pro střední hodnotu a rozptyl $\hat{\alpha}$ platí

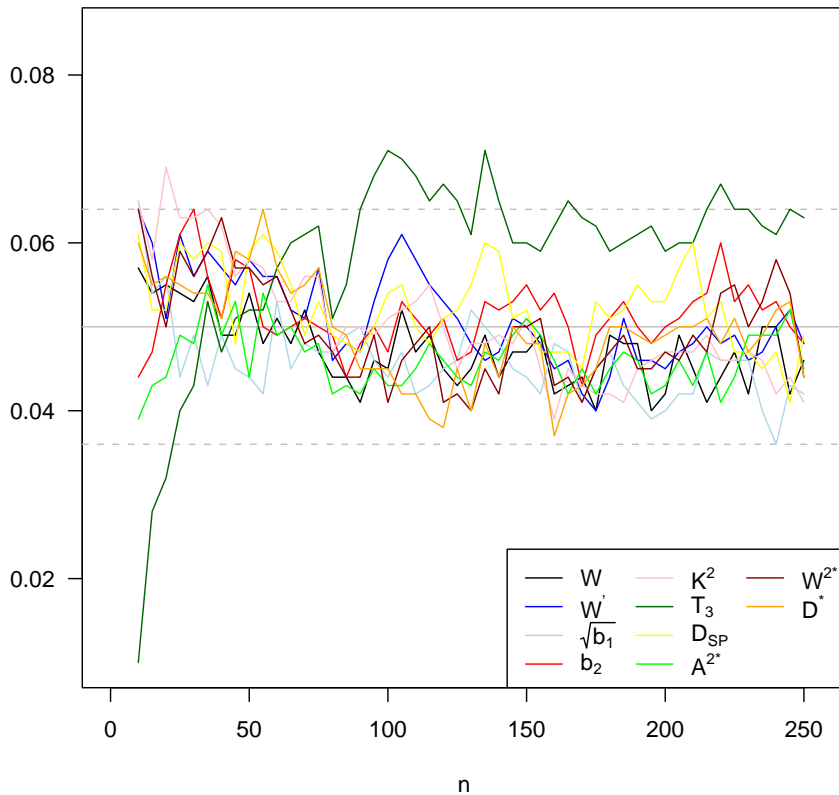
$$\begin{aligned} E \hat{\alpha} &= \frac{1}{1000} \sum_{i=1}^{1000} E X_i = \alpha \\ \text{var } \hat{\alpha} &= \frac{1}{1000^2} \sum_{i=1}^{1000} \text{var } X_i = \frac{\alpha(1 - \alpha)}{1000}. \end{aligned}$$

Pro námi použitou hladinu $\alpha = 0,05$ můžeme určit kritický obor s hraničními hodnotami 0,036 a 0,064. Padne-li tedy odhad $\hat{\alpha}$ mimo tento interval, můžeme říci, že se na hladině 5 % statisticky významně liší od požadované hladiny $\alpha = 0,05$.

Graf na obrázku 4.1 znázorňuje odhad hladiny testu $\hat{\alpha}$ v závislosti na rozsahu výběru n pro jednotlivé testy aplikované přímo na vektor chyb \mathbf{e} . Z něj je patrné, že pro vektor chyb \mathbf{e} dodržují všechny testy kromě T_3 a K^2 předepsanou hladinu. Zatímco test K^2 nedodržel hladinu 5 % pouze pro rozsah výběru $n = 20$ ($\hat{\alpha} = 0,069$), test T_3 vykazuje četnější nedodržování hladiny. Pro rozsahy výběru $n < 25$ je $\hat{\alpha}$ výrazně nižší než požadovaných 5 % a pro zhruba třetinu rozsahů výběru $n > 100$ již naopak odhad hladiny testu $\hat{\alpha}$ vychází vyšší než horní kritická mez.

Pokud aplikujeme jednotlivé testy na vektory reziduí místo na vektor chyb \mathbf{e} , odhady hladiny testů vyjdou odlišně. V případě modelu regresní přímky však tento rozdíl pro rezidua \mathbf{u} a \mathbf{v} není příliš výrazný. Použití nekorelovaných reziduí \mathbf{n} vede ke snížení odhadu hladiny testů A^{2*} a W^{2*} tak, že se $\hat{\alpha}$ pro většinu rozsahů výběru $80 < n < 175$ výrazně liší od 5 %. Odhad hladiny testu T_3 naopak pro většinu rozsahů $n > 75$ vychází vyšší než horní kritická mez pro námi požadovanou hladinu testu $\alpha = 5 \%$.

Největší vliv na odhad hladiny testů má použití studentizovaných reziduí. Pro malé rozsahy výběrů ($n < 50$) nedodrží předepsanou hladinu žádný



Obrázek 4.1: Hladina testů aplikovaných na vektor \mathbf{e}

z testů. Odhady hladiny testů W' , $\sqrt{b_1}$ a K^2 pro $n = 10$ dosahují dokonce 18 % resp. 21 %. S rostoucím rozsahem výběru se situace zlepšuje, takže pro $n > 50$ již můžeme $\hat{\alpha}$ pro testy A^{2*} , D^* , T_3 a W^{2*} považovat za vyhovující. Ostatním testy potřebují ještě větší rozsahy výběru - pro W , $\sqrt{b_1}$, b_2 je nutný rozsah výběru alespoň $n = 110$. Pro $n > 150$ už dodržují hladinu všechny testy kromě testu T_3 . Pro něj i při takto velkých rozsazích výběru dosahuje $\hat{\alpha}$ hodnoty 8–10 % místo požadovaných $\alpha = 5\%$.

Grafy znázorňující odhady hladin testů při použití jednotlivých reziduí v modelu regresní přímky jsou pro přehlednost znázorněny v příloze na obrázku A.7.

V modelu kvadratické závislosti se použití reziduí \mathbf{u} a \mathbf{v} místo vektoru chyb \mathbf{e} projeví nejvíce na odhadu hladiny testu T_3 . Pro rozsahy výběru $n > 95$ už překračuje horní kritickou mez pro $\alpha = 5\%$. Odhady hladin ostatních testů

se až na drobné výjimky pro rozsah výběru $n = 10$ neliší od požadované hladiny $\alpha = 5\%$.

Při použití nekorelovaných reziduí \mathbf{n} , dochází k častějšímu nedodržování požadované hladiny než v modelu regresní přímky. Kromě testu T_3 , který překračuje horní kritickou mez pro všechny sledované rozsahy $n > 80$, je patrně častější překročení horní kritické meze pro hladinu $\alpha = 5\%$ v případě testů D_{SP} a W' .

Použití studentizovaných reziduí vede k ještě výraznějšímu nedodržování hladiny pro malé rozsahy výběrů než v případě modelu regresní přímky. Testy A^{2*} , D^* , D_{SP} a W^{2*} dodržují požadovanou hladinu až pro $n > 60$, testy W , $\sqrt{b_1}$, b_2 dokonce až pro $n > 120$. Testům W' a K^2 se podařilo dodržet hladinu pouze pro několik málo velkých rozsahů výběru a odhad hladiny testu T_3 se opět pro všechny rozsahy výběru pohybuje mezi 8 a 12%. Grafy srovnávající $\hat{\alpha}$ pro jednotlivé testy při použití reziduí v modelu kvadratické závislosti jsou v příloze na obrázku A.8.

V případě modelu regrese se třemi regresory a absolutním členem je zajímavá reakce odhadu hladiny testu T_3 na použití některých reziduí. Pro rezidua \mathbf{u} a \mathbf{v} tento test pro větší rozsahy výběru lépe dodržuje požadovanou hladinu $\alpha = 5\%$ než v případě použití vektoru chyb \mathbf{e} . Pro tato rezidua ostatní testy (kromě několika výjimek pro rozsah $n = 10$) dodržují hladinu dobře. V případě nekorelovaných reziduí je odhad hladiny testů A^{2*} a W^{2*} spíše podhodnocen. Naopak $\hat{\alpha}$ testu D_{SP} překračuje horní kritickou mez pro $n \leq 30$. Při použití studentizovaných reziduí \mathbf{v}^* místo vektoru chyb \mathbf{e} opět dochází k výraznému nedodržení hladiny testů. Nejlépe se s tím vyrovnávají testy A^{2*} , D^* , D_{SP} a W^{2*} , které dodržují předepsanou hladinu již pro rozsah výběru $n > 50$. O ostatních testech můžeme říci, že dodržují předepsanou hladinu až pro rozsahy výběru $n > 110$ (W , $\sqrt{b_1}$), $n > 130$ (pro b_2), resp. $n > 145$ (pro K^2). Shapirův–Franciův test má problémy s dodržováním hladiny i pro velké rozsahy výběru a test T_3 překračuje horní kritickou mez pro všechny rozsahy výběru n . Pro lepší představu viz grafy na obrázku A.9.

Při použití reziduí \mathbf{u} , \mathbf{v} místo vektoru chyb \mathbf{e} v modelu jednoduchého třídění nedošlo k téměř žádným změnám v odhadech hladin jednotlivých testů. Pro studentizovaná rezidua se ovšem odhad hladiny testů opět výrazně zhoršil. Pro malé rozsahy výběrů n nedodrží předepsanou hladinu žádný z testů. Nejlépe se s použitím \mathbf{v}^* vyrovnává test D_{SP} , který dodržuje předepsanou hladinu $\alpha = 5\%$ již pro $n > 40$. O něco větší rozsah výběru potřebují testy A^{2*} , W^{2*} a D^* , které dodržují hladinu pro rozsahy $n > 60$. Ostatní testy začínají dodržovat hladinu až pro velké rozsahy výběru a test T_3 hladinu nedodržuje vůbec (pro všechny rozsahy výběru se $\hat{\alpha}$ pohybuje mezi 8 a 12%). Viz grafy v obrázku A.10.

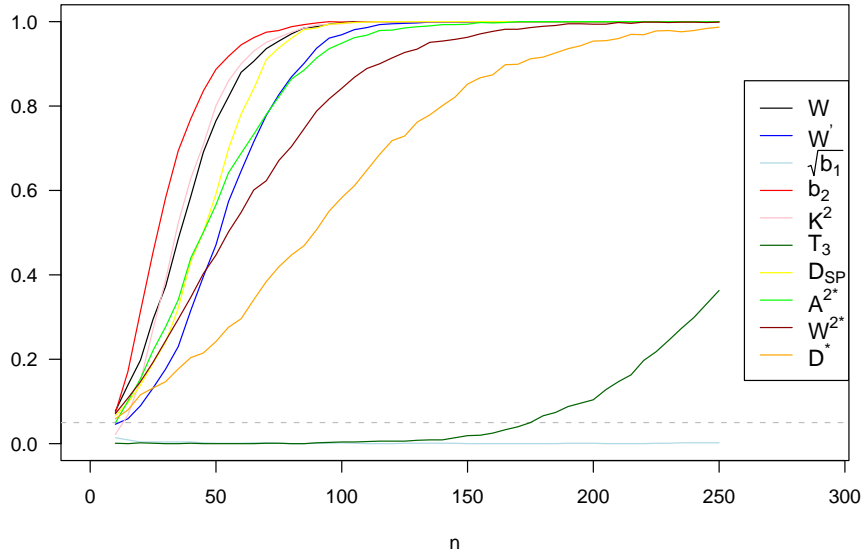
4.2 Srovnání síly testů

Sílu testů proti jednotlivým alternativám jsme odhadovali tak, že jsme vždy vzali 1000 výběrů z příslušného rozdělení a sledovali počet zamítnutí. Veškeré výsledky jsou pak uvedeny v procentech.

Pro rozdělení $R(-\sqrt{3}, \sqrt{3})$ mají nejmenší sílu testy $\sqrt{b_1}$ a T_3 . Test $\sqrt{b_1}$ má téměř nulovou sílu nezávisle na rozsahu výběru, protože je založen pouze na šikmosti, kterou má rovnoměrné rozdělení shodnou s normálním rozdělením. Síla testu T_3 z téměř nuly pro $n < 150$ roste až na 36 % pro $n = 250$. Naopak nejlépe dopadl test b_2 založený na špičatosti rozdělení. Jeho síla rychle roste ze 7 % pro $n = 10$ až na 100 % pro rozsah výběru $n = 100$. Z testu b_2 je odvozen i test K^2 , u kterého síla roste jen o málo pomaleji (z 8 % pro $n = 10$ na 100 % pro $n > 100$). Podobně vypadá i průběh silofunkce pro test W . Ten vykazuje dokonce vyšší sílu než K^2 pro rozsah výběru menší než 30. Síla testů A^{2*} a D_{SP} roste shodně z 5–7 % pro $n = 10$ na 50 % pro $n = 45$. Pro vyšší rozsah výběru roste síla testu D_{SP} rychleji - dosahuje 100 % už pro $n > 100$. Test A^{2*} dosahuje síly 100 % až pro $n > 150$. Cramerův-von Misesův test W^{2*} má pro $n < 45$ o málo (přibližně o 6 procentních bodů) vyšší sílu než W . Pro $n > 45$ už ale síla W^{2*} roste pomaleji a 100 % dosahuje až pro rozsah výběru $n > 200$. Nejslabší z testů, které vykazují alespoň nějakou sílu, je Lillieforsova modifikace Kolmogorovova-Smirnovova testu. Síla testu D^* roste pomalu z 6 % pro $n = 10$ na 99 % pro $n = 250$. Popsané výsledky jsou lépe patrné z grafu na obrázku 4.2.

V případě rozdělení $Exp(2)$ vychází u všech testů poměrně velká síla, protože se toto rozdělení díky své asymetrii výrazně liší od normálního rozdělení. Mezi testy $W, W', \sqrt{b_1}, D_{SP}, W^{2*}$ a A^{2*} nejsou větší rozdíly v síle. Ta rychle roste z 30–40 % pro $n = 10$ na 100 % pro $n = 50$. Pouze v případě Cramerova-von Misesova testu klesá síla pro $n > 165$ ze 100 % až na 88 % pro rozsah výběru $n = 250$. Trochu slabší jsou testy K^2 a D^* , které pro $n = 10$ dosahují síly přibližně 29 %. Pro takto malé rozsahy výběru je výrazně slabší test T_3 (síla pouze 9 %), ale pro $n = 70$ už všechny tři posledně zmiňované testy dosahují síly 100 %. Nejhůře dopadl test b_2 založený pouze na špičatosti. V jeho případě roste síla v závislosti na rozsahu výběru spíše pozvolně z 20 % pro $n = 10$ a 100 % dosahuje až pro $n = 250$. Pro lepší představu je zde graf na obrázku 4.3.

Pro rozdělení t_5 je celkově síla testů nižší než v případě rozdělení $Exp(2)$. Jako nejsilnější se jeví skupina testů W', b_2, W, T_3 a K^2 . Jejich síla pomalu roste z 11–15 % (kromě testu T_3 , který dosahuje pouze 2 %) na 85–90 % pro $n = 250$. O něco slabší jsou testy A^{2*} a W^{2*} s dosaženou silou 7–9 % pro $n = 10$, která pro $n = 250$ vzroste až na 76–81 %. Ještě o něco pomalejší je růst síly u testů D_{SP} a D^* (z 9 % na 56–62 %). Zajímavý růst síly vykazuje test

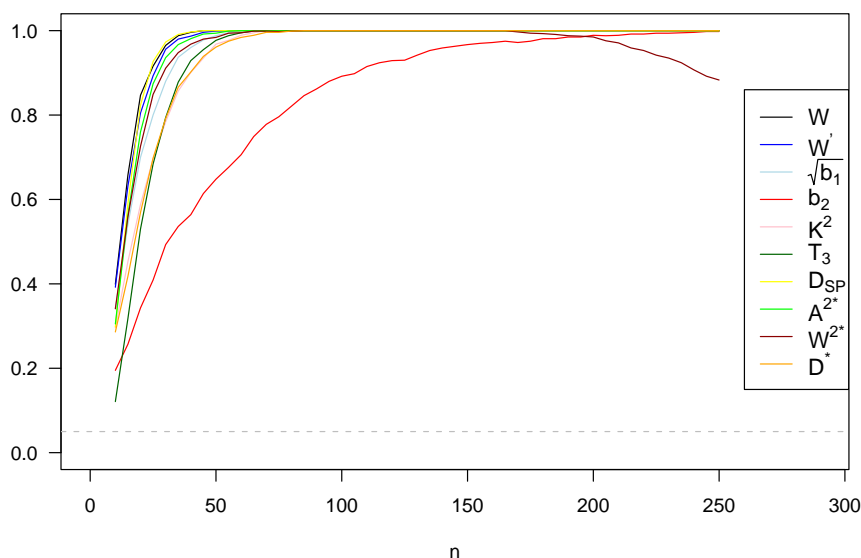


Obrázek 4.2: Síla testů aplikovaných na vektor \mathbf{e} pocházející z rozdělení $R(-\sqrt{3}, \sqrt{3})$

$\sqrt{b_1}$. Pro rozsahy výběru $n < 40$ je jeho síla srovnatelná s testy W' , b_2 , W , T_3 a K^2 . Pro větší rozsahy výběrů se růst síly zpomaluje, takže pro $n = 250$ dosahuje síly pouze 50%. Viz graf na obrázku 4.4.

Pro dvojitě exponenciální rozdělení je nejsilnější test W' . Jeho síla roste z 19% pro $n = 10$ až na 99% pro rozsah výběru $n = 250$. Pro malé rozsahy výběru $n < 35$ mu může konkurovat souhrnný test K^2 , jehož síla ale pro větší rozsahy výběrů roste pomaleji a pro $n = 250$ dosahuje 97%. Síla testů A^{2*} , W^{2*} , b_2 a W roste shodně z 14–16% na 34% pro $n = 30$. Pro větší rozsahy výběrů jsou mezi nimi patrné menší rozdíly v síle (maximálně však 6 procentních bodů), ale pro $n = 250$ už shodně dosahují síly 99%. Nejmenší sílu vykazuje pro malé rozsahy výběrů test T_3 (pouhá 4% pro $n = 10$). Pro $n > 25$ už však dosahuje obdobných výsledků jako test D^* . Oba testy nakonec pro $n = 250$ dosahují síly přibližně 96%. Test D_{SP} je poměrně slabý - jeho síla roste pomaleji než u ostatních testů a pro $n = 250$ dosahuje 90%. Nejhuře dopadl test $\sqrt{b_1}$ založený pouze na šikmosti, kterou má dvojitě exponenciální rozdělení shodnou s normálním rozdělením. Pro něj síla velmi pomalu roste z 18% pouze na 46,5%. Lépe je vše patrné z grafu na obrázku 4.5.

Pokud vektor chyb \mathbf{e} pochází ze směsi dvou normálních rozdělení s nulovou střední hodnotou a různými rozptyly, je síla všech testů poměrně nízká.



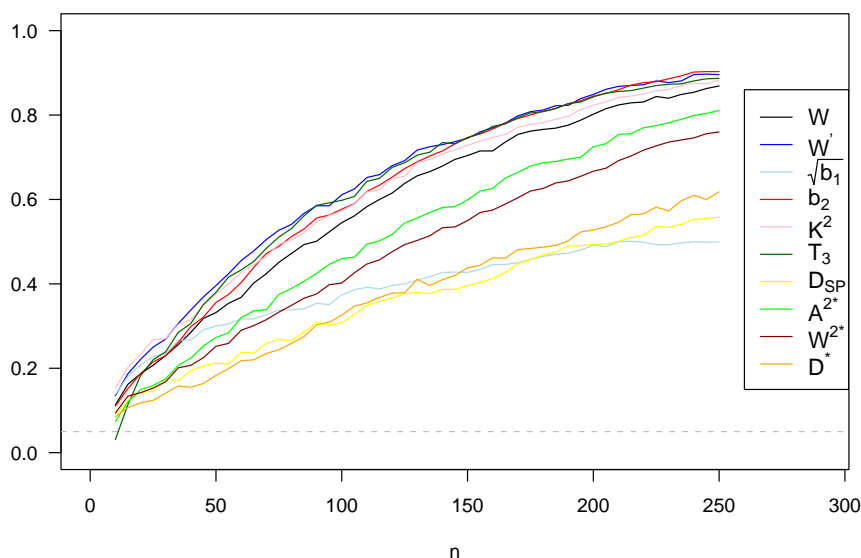
Obrázek 4.3: Síla testů aplikovaných na vektor \mathbf{e} pocházející z rozdělení $Exp(2)$

Nejlépe dopadl test W' jehož síla roste z 11 % na 60 % pro $n = 250$. Test T_3 pro rozsah výběru $n = 10$ nabývá síly pouze 2 %, ale s rostoucím n se přibližuje skupině testů A^{2*} , W^{2*} , b_2 , K^2 a W , jejichž síla roste ze 7–11 % na 49–56 %. Výrazně slabší jsou testy D_{SP} a D^* , jejichž síla roste z 8 % na 22 % resp. 39 %. Nejhůře dopadl test $\sqrt{b_1}$ založený na šikmosti, jehož síla se pro všechny uvažované rozsahy výběru n pohybuje mezi 10 a 17 %. Viz graf na obrázku 4.6.

4.2.1 Model regresní přímky

V případě vektoru chyb \mathbf{e} pocházejícího z rozdělení $R(-\sqrt{3}, \sqrt{3})$ není v síle testů rozdíl při použití reziduí \mathbf{u} , \mathbf{v} a \mathbf{v}^* . Oproti \mathbf{e} však tato rezidua nepatrně snižují sílu testů $\sqrt{b_1}$, b_2 , K^2 , T_3 , D^* , W^{2*} a A^{2*} . U testů W , W' a D_{SP} je rozdíl při použití těchto reziduí až 15 procentních bodů, ale pro $n > 110$ se rozdíly stírají. K nejvýraznějšímu poklesu síly testů dochází při aplikaci testů na vektor nekorelovaných reziduí \mathbf{n} . Síla testů v tomto případě klesá až o 40 procentních bodů. Viz graf na obr. A.11.

Pro rozdělení $Exp(2)$ není patrný výraznější pokles síly pro rezidua \mathbf{u} , \mathbf{v} a \mathbf{v}^* . Zajímavý je však vliv použití nekorelovaných reziduí \mathbf{n} . U všech testů

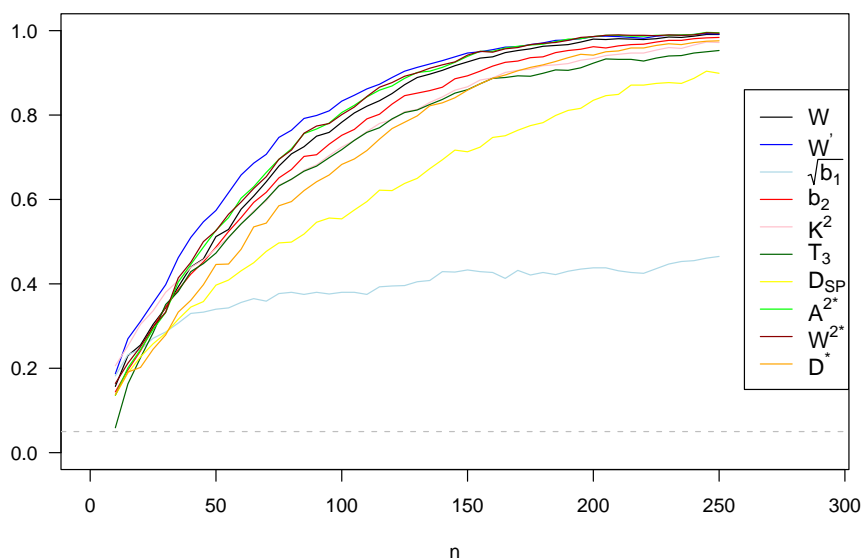


Obrázek 4.4: Síla testů aplikovaných na vektor \mathbf{e} pocházející z rozdělení t_5

dochází ke snížení síly pro malé rozsahy výběrů n a to až o 40 procentních bodů. Pro test W^{2*} je však pro rozsahy výběrů $n > 170$ síla testů aplikovaných na \mathbf{n} vyšší než při použití \mathbf{e} . Viz graf A.12.

Pokud vektor \mathbf{e} pochází z rozdělení t_5 , klesá síla testů při použití reziduí jen zanedbatelně (o max. 4 procentní body pro rezidua \mathbf{u}, \mathbf{v} a max. o 8 procentních bodů pro rezidua \mathbf{n}). Pouze pro studentizovaná rezidua dochází ke zvýšení síly testů oproti použití chybového vektoru \mathbf{e} o 12–15 procentních bodů pro $n = 10$. Tento rozdíl se s rostoucím n snižuje až na 2–5 procentních bodů pro $n = 250$. Viz graf na obrázku A.13. Z grafu na obrázku A.14 je vidět, že velmi podobně na použití reziduí reagují testy i v případě vektoru \mathbf{e} pocházejícího ze smíšeného normálního rozdělení.

V případě dvojitě exponenciálního rozdělení je snížení síly v důsledku použití reziduí \mathbf{u}, \mathbf{v} opět zanedbatelné, ale více se projeví použití nekorelovaných reziduí \mathbf{n} . Pro všechny testy je toto snížení síly maximálně o 17 procentních bodů s tím, že citlivější jsou testy A^{2*} , W^{2*} a D^* . Při použití studentizovaných reziduí \mathbf{v}^* dochází opět ke zvýšení síly testů oproti použití chybového vektoru \mathbf{e} a to o 13–17 procentních bodů pro $n = 10$. Tyto rozdíly se s rostoucím n stírají, takže pro $n > 110$ jsou již menší než 5 procentních bodů. Lépe je vše patrné z grafu na obrázku A.15.



Obrázek 4.5: Síla testů aplikovaných na vektor \mathbf{e} pocházející z rozdělení $DExp(1, \frac{1}{\sqrt{2}})$

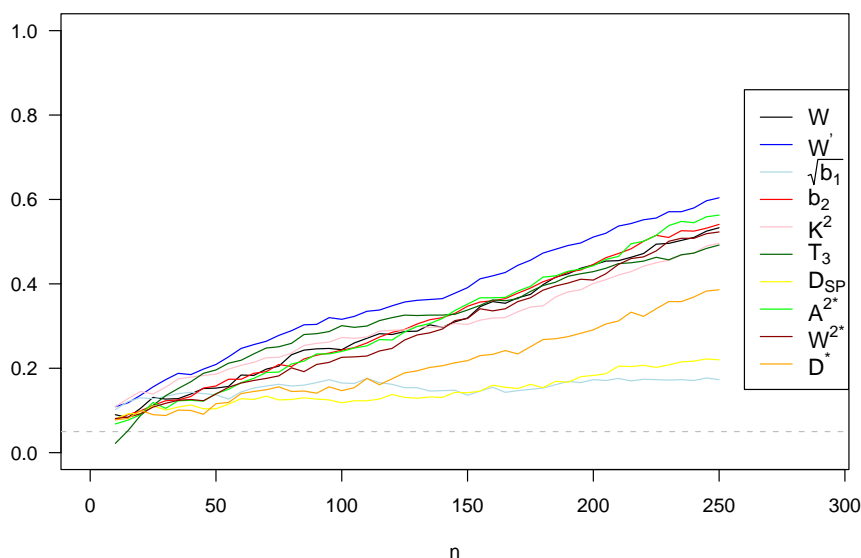
4.2.2 Model kvadratické závislosti

V případě rozdělení $R(-\sqrt{3}, \sqrt{3})$ mají všechny testy nejvyšší sílu při použití chybového vektoru \mathbf{e} . Ve svém vlivu na sílu testů se od sebe rezidua $\mathbf{u}, \mathbf{v}, \mathbf{v}^*$ navzájem neliší - všechna sílu testů snižují stejně. Pro testy W^{2*}, A^{2*} se jedná o maximální rozdíly 11–15 procentních bodů, pro test D^* méně než 8 procentních bodů a pro ostatní testy kromě $\sqrt{b_1}$ a T_3 je tento rozdíl 19–25 procentních bodů. S rostoucím n se výše zmíněný pokles síly zmenšuje.

Při použití nekorelovaných reziduí \mathbf{n} dochází k výrazně většímu poklesu síly a to až o 50 procentních bodů. Tyto rozdíly se srovnávají teprve pro větší rozsahy výběru než v případě reziduí $\mathbf{u}, \mathbf{v}, \mathbf{v}^*$. Pro lepší představu jsou příslušné grafy v příloze na obrázku A.16.

Pro rozdělení $Exp(2)$ se vliv použití reziduí místo chybového vektoru \mathbf{e} projevuje podobně jako v případě rovnoměrného rozdělení. Pro rezidua $\mathbf{u}, \mathbf{v}, \mathbf{v}^*$ dochází v případě testů $W, W', \sqrt{b_1}, D_{SP}, W^{2*}, D^*$ a A^{2*} k výraznějšímu poklesu síly o 18–30 procentních bodů pro rozsahy výběru $n < 30$. S rostoucím n se pak tyto rozdíly rychle stírají. U zbylých testů jsou rozdíly v síle menší.

Při použití nekorelovaných reziduí \mathbf{n} dochází opět k většímu poklesu síly



Obrázek 4.6: Síla testů aplikovaných na vektor \mathbf{e} pocházející ze smíšeného normálního rozdělení

než v případě ostatních reziduí. Pro testy W , W' a D_{SP} dokonce až o 50 procentních bodů. U všech testů se tyto rozdíly stávají zanedbatelnými až pro podstatně větší rozsahy výběrů než v případě ostatních reziduí. Jedinou výjimku tvoří opět test W^{2*} , který naopak pro $n > 165$ vykazuje vyšší sílu při použití reziduí než při použití chybového vektoru \mathbf{e} . Pro lepší představu viz grafy na obrázku A.17.

V případě rozdělení t_5 není téměř rozdíl mezi aplikací testů na vektor chyb \mathbf{e} a na rezidua \mathbf{u} , \mathbf{v} . Při použití nekorelovaných reziduí \mathbf{n} klesá síla pouze nepatrně (o méně než 10 procentních bodů). Pro studentizovaná rezidua \mathbf{v}^* opět dochází ke zvýšení síly jednotlivých testů a to až o 8–13 procentních bodů s tím, že se tyto rozdíly s rostoucím n zmenšují. Viz graf na obrázku A.18. Podobné výsledky dostáváme i pro vektor chyb pocházející ze směsi dvou normálních rozdělení. Příslušné grafy jsou v příloze na obrázku A.19.

Pro vektor chyb \mathbf{e} pocházející z rozdělení $DExp(1, \frac{1}{\sqrt{2}})$ je snížení síly testů v důsledku použití reziduí \mathbf{u} , \mathbf{v} poměrně nízké - do 10 procentních bodů. Použijeme-li nekorelovaná rezidua \mathbf{n} , nejvíce poklesne dosažená síla u testů W^{2*} , D^* , A^{2*} a to až o 21 procentní bod. Pro ostatní testy je pokles menší (12–16 procentních bodů). Při aplikaci testů na studentizovaná rezidua místo na vektor chyb \mathbf{e} dochází k mírnému zvýšení síly (maximálně však

o 14 procentních bodů). Pro lepší představu je v příloze obrázek A.20.

4.2.3 Model regrese se třemi regresory

V modelu lineární regrese s absolutním členem a třemi regresory se použití reziduí $\mathbf{u}, \mathbf{v}, \mathbf{v}^*$ místo vektoru \mathbf{e} , který pochází z rozdělení $R(-\sqrt{3}, \sqrt{3})$, projeví nejméně při použití Lillieforsova testu D^* a testu W^{2*} (pokles síly o maximálně 10 resp. 19 procentních bodů). U ostatních testů dochází k většímu poklesu síly - a to až o 40 procentních bodů proti aplikaci testů přímo na vektor chyb \mathbf{e} . U všech testů se však tyto rozdíly pro větší rozsahy výběru n stírají. Při použití nekorelovaných reziduí \mathbf{n} je pokles síly ještě výraznější - dokonce až o 60 procentních bodů. Pro lepší představu jsou zde grafy na obrázku A.21.

Pro vektor \mathbf{e} pocházející z rozdělení $Exp(2)$ opět nedochází k téměř žádnému snížení síly testů $\sqrt{b_1}, K^2$ a T_3 při použití reziduí $\mathbf{u}, \mathbf{v}, \mathbf{v}^*$. U ostatních testů dochází k výraznějšímu snížení síly (až o 30 procentních bodů v případě testu D_{SP}). Pouze pro test založený na špičatosti b_2 dochází k jejímu mírnému zvýšení. Pokud testy aplikujeme na nekorelovaná rezidua \mathbf{n} , dojde k výraznému poklesu jejich síly (až o 62 procentních bodů). Pro všechny testy se zmíněné rozdíly v síle s rostoucím n zmenšují (s výjimkou testu W^{2*} , pro který naopak síla pro $n > 170$ při použití reziduí roste). Viz grafy na obrázku A.22.

V případě, že vektor chyb \mathbf{e} pochází z rozdělení t_5 , není téměř rozdíl v síle testů aplikovaných na rezidua \mathbf{u} a \mathbf{v} , která se navíc pouze minimálně liší od síly testů pro vektor \mathbf{e} . Použití nekorelovaných reziduí vede k mírnému snížení síly testů (maximálně však o 12 procentních bodů). Studentizovaná rezidua sílu testů naopak zvyšují. Pro lepší představu jsou v příloze grafy na obrázku A.23.

Pro směs dvou normálních rozdělení je pokles síly testů aplikovaných na rezidua $\mathbf{u}, \mathbf{v}, \mathbf{n}$ malý nezávisle na rozsahu výběru. Pokud použijeme studentizovaná rezidua \mathbf{v}^* , dochází naopak ke zvýšení síly (pro malé rozsahy výběru až o 13 procentních bodů), přičemž se tento rozdíl s rostoucím n zmenšuje. Graficky jsou tyto výsledky znázorněny v příloze v grafech na obrázku A.24.

Použijeme-li místo vektoru \mathbf{e} pocházejícího z dvojité exponenciálního rozdělení rezidua \mathbf{u}, \mathbf{v} , dojde k mírnému snížení síly testů (maximálně však o 13 procentních bodů). Naopak použití studentizovaných reziduí vede ke zvýšení síly (pro malé rozsahy až o 15 procentních bodů), které se s rostoucím n zmenšuje. Výrazněji se projeví použití nekorelovaných reziduí \mathbf{n} . Pro ně dochází ke snížení síly testů až o 26 procentních bodů (kromě testu $\sqrt{b_1}$, pro který tento rozdíl nepřesáhne 12 procentních bodů). Viz grafy na obrázku A.25.

4.2.4 Model jednoduchého třídění

Jak již bylo zmíněno výše, v modelu jednoduchého třídění nepoužíváme nekorelovaná rezidua, protože v něm nemají smysluplnou interpretaci.

Pokud vektor \mathbf{e} pochází z rovnoměrného rozdělení $R(-\sqrt{3}, \sqrt{3})$, není téměř rozdíl v síle testů při použití reziduí $\mathbf{u}, \mathbf{v}, \mathbf{v}^*$. Všechny podobně snižují sílu testů a to až o 21 procentní bod. Rozdíly v síle se s rostoucím n snižují, takže pro $n > 110$ (v případě testu D^* až pro $n > 170$) už jsou menší než 5 procentních bodů. Viz grafy na obrázku A.26.

V případě exponenciálního rozdělení $Exp(2)$ jsou patrné rozdíly při použití reziduí pro malé rozsahy výběru n . Rezidua \mathbf{u}, \mathbf{v} tak snižují sílu testů až o 20 procentních bodů, ale pro $n > 40$ už je toto snížení pro všechny testy menší než 5 procentních bodů. Použití studentizovaných reziduí \mathbf{v}^* vede naopak ke zvýšení síly v porovnání s aplikací testů na vektor chyb \mathbf{e} . Tyto rozdíly jsou maximálně 16 procentních bodů a s rostoucím n rychle klesají, takže pro $n > 35$ jsou již pro všechny testy kromě b_2 menší než 5 procentních bodů. Rozdíl v síle pro test b_2 klesá pod zmiňovanou mez až pro rozsahy výběru $n > 70$. Příslušné grafy jsou v příloze na obrázku A.27.

Pro vektor \mathbf{e} pocházející z rovnoměrného rozdělení snižuje použití reziduí \mathbf{u}, \mathbf{v} dosaženou sílu jen minimálně. Pokud však testy aplikujeme na studentizovaná rezidua \mathbf{v}^* , dojde ke zvýšení síly testů. Rozdíl činí až 16 procentních bodů a s rostoucím n se pomalu snižuje. Grafy znázorňující výše popsané vlivy použití reziduí jsou v příloze na obrázku A.28. Z grafů na obrázku A.29 je vidět, že se velmi podobně projeví použití reziduí i v případě vektoru \mathbf{e} , který je směsí dvou normálních rozdělení.

Také pro dvojitě exponenciální rozdělení není patrný téměř žádný pokles síly při použití reziduí \mathbf{u}, \mathbf{v} místo vektoru chyb \mathbf{e} . Pro studentizovaná rezidua pozorujeme opět mírné zvýšení dosažené síly (maximálně o 17 procentních bodů) s tím, že se tento rozdíl s rostoucím rozsahem výběru n postupně stírá. Pro lepší představu viz grafy na obrázku A.30.

Závěr

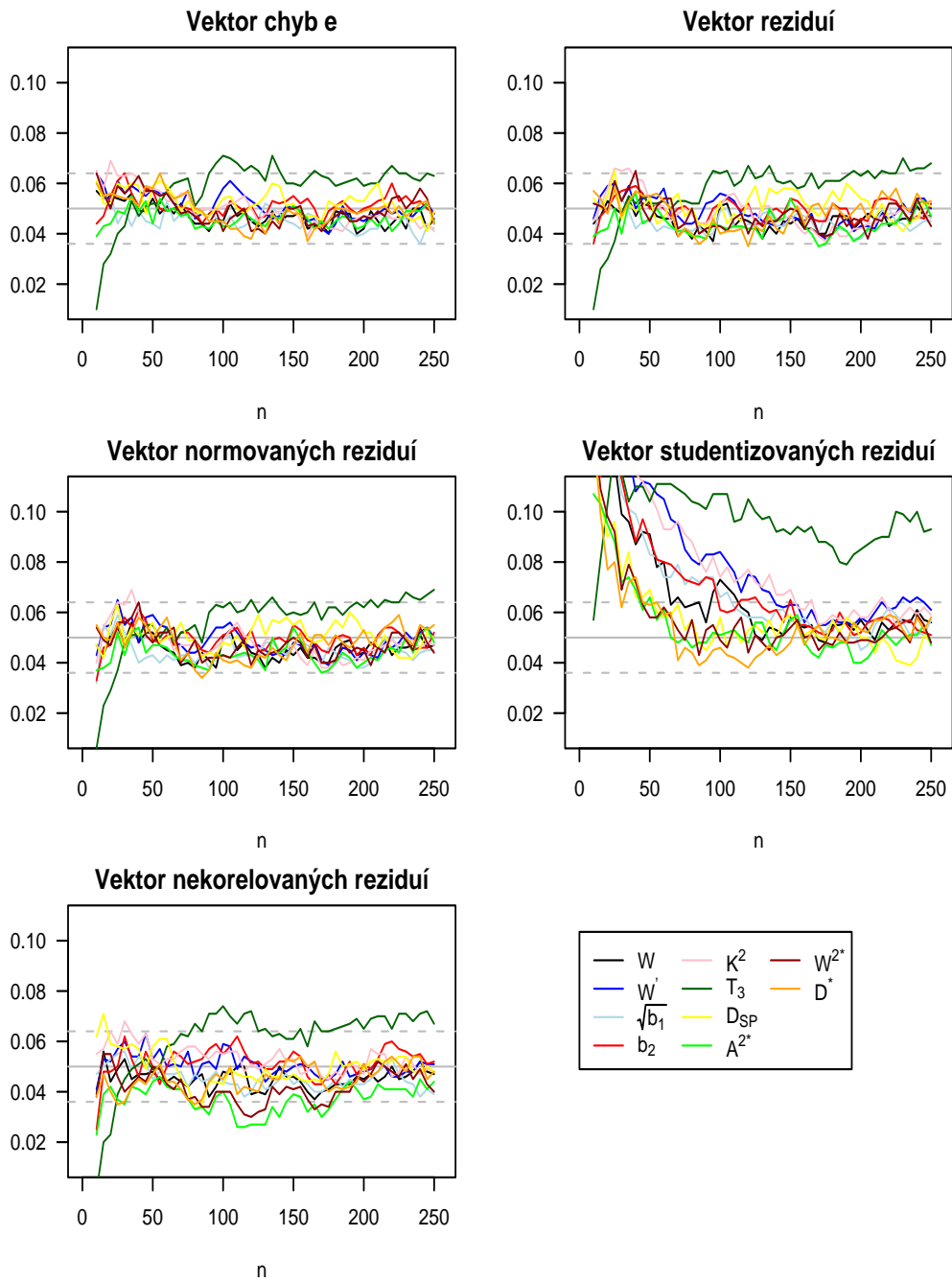
Pro testování hypotéz v lineárním modelu je podstatný předpoklad, že složky chybového vektoru \mathbf{e} , které tvoří náhodnou složku vysvětlované veličiny, mají normální rozdělení. Protože chybový vektor jako takový nikdy nepozorujeme, je nutné předpoklad normality ověřovat pomocí některého typu reziduí. K tomu máme, kromě subjektivního diagnostického prostředku, kterým je diagram normality, k dispozici i objektivní statistické testy.

V této simulační studii jsme srovnávali vliv použití různých reziduí na dodržování hladiny testu. Ukázalo se, že běžně používaná rezidua \mathbf{u} a z nich odvozená normovaná rezidua \mathbf{v} téměř neovlivňují hladinu srovnávaných testů. Nekorelovaná rezidua \mathbf{n} v závislosti na matici modelu mírně zhoršují dodržování požadované hladiny testu. Použití studentizovaných reziduí \mathbf{v}^* se ukazuje jako nevhodné, protože vede k překračování požadované hladiny $\alpha = 5\%$ (pro malé rozsahy výběru $\hat{\alpha}$ vychází až 20%). To je nejspíše způsobeno tím, že pokud vektor \mathbf{e} pochází z normálního rozdělení, mají studentizovaná rezidua rozdělení t_{n-r-1} a navíc nejsou nezávislá.

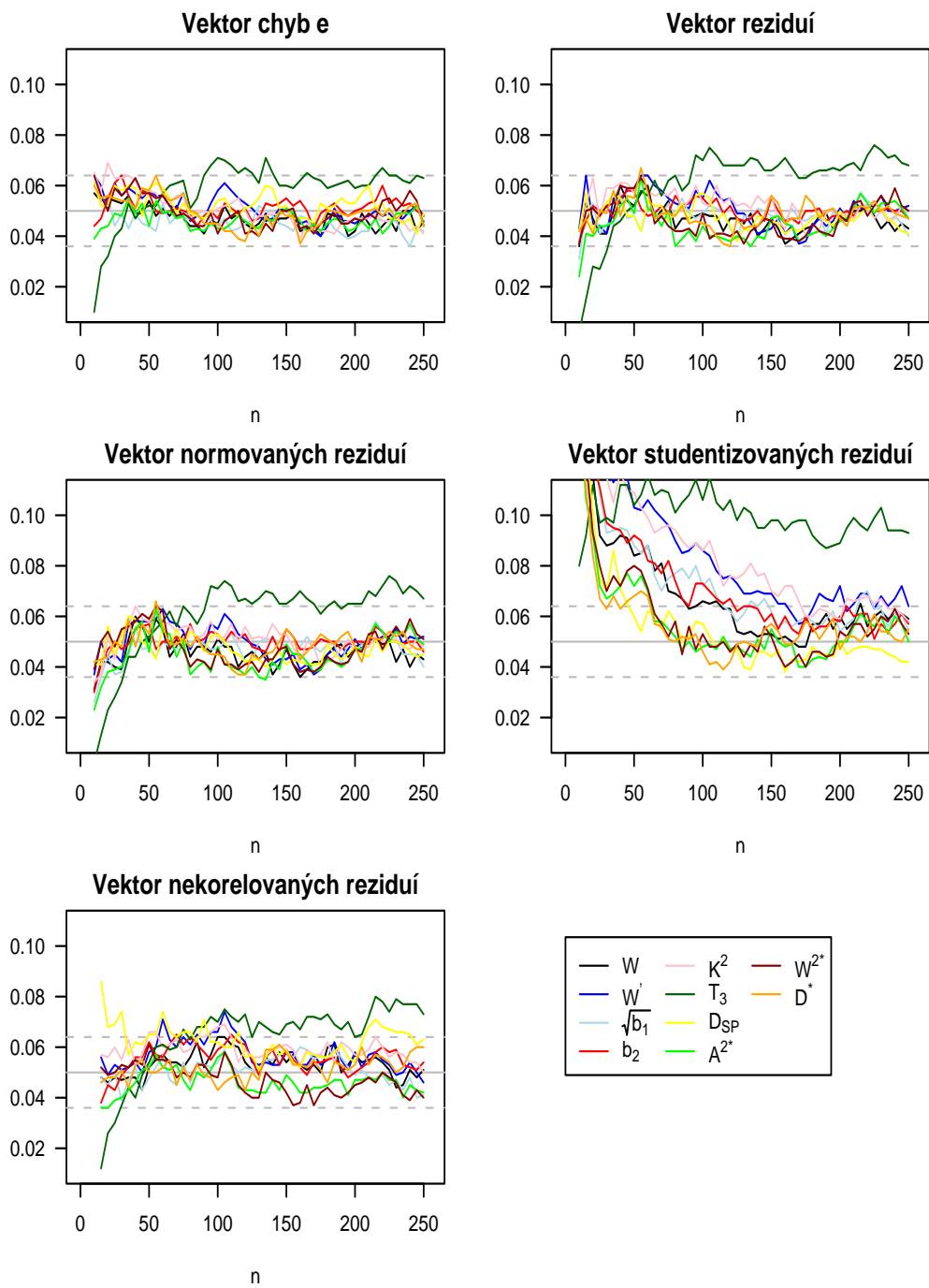
Při srovnávání síly testů proti jednotlivým alternativám se ukazuje, že poměrně neznámý Michaelův test D_{SP} se řadí spíše ke slabším testům a navíc na použití nekorelovaných reziduí reaguje více než ostatní testy. Druhý poměrně neznámý test T_3 vykazuje větší sílu než D_{SP} pro většinu sledovaných alternativ, ale např. pro rovnoměrné rozdělení zcela selhává. Navíc tento test nedodržuje příliš dobře předepsanou hladinu. Z ostatních testů velmi dobře dopadly známé testy Shapirův-Wilkův (W), Shapirův-Franciov (W') a souhrnný test K^2 . Naopak se příliš neosvědčily testy zaměřené pouze na šikmost ($\sqrt{b_2}$) či špičatost (b_2), jejichž úzká specializace se hodí spíše v případě, kdy předpokládáme odchylku od normality v šikmosti či špičatosti.

Ukázalo se, že jednotlivé modely se ve svém vlivu na dodržování hladiny a sílu testů příliš neliší.

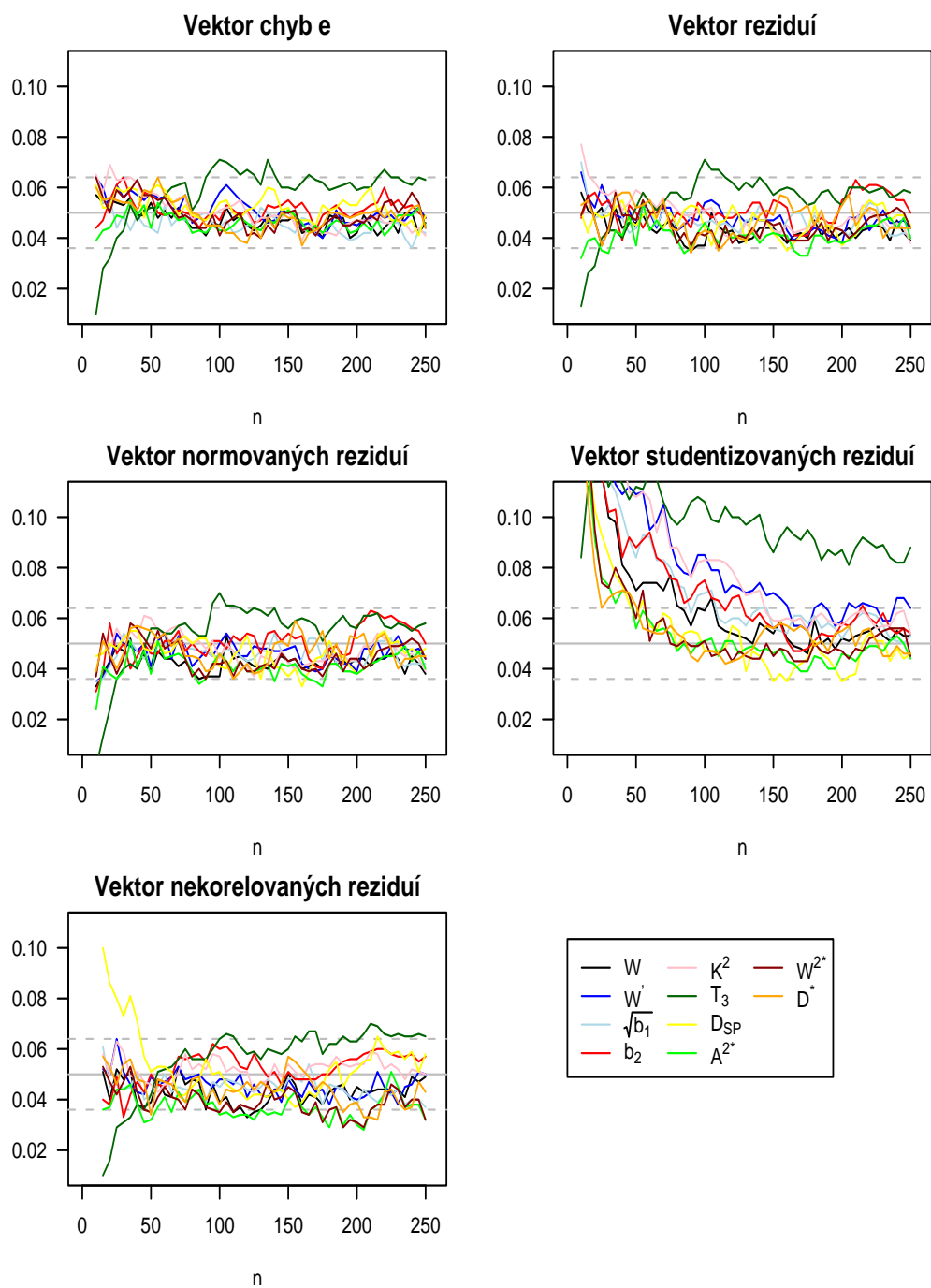
A Obrázky



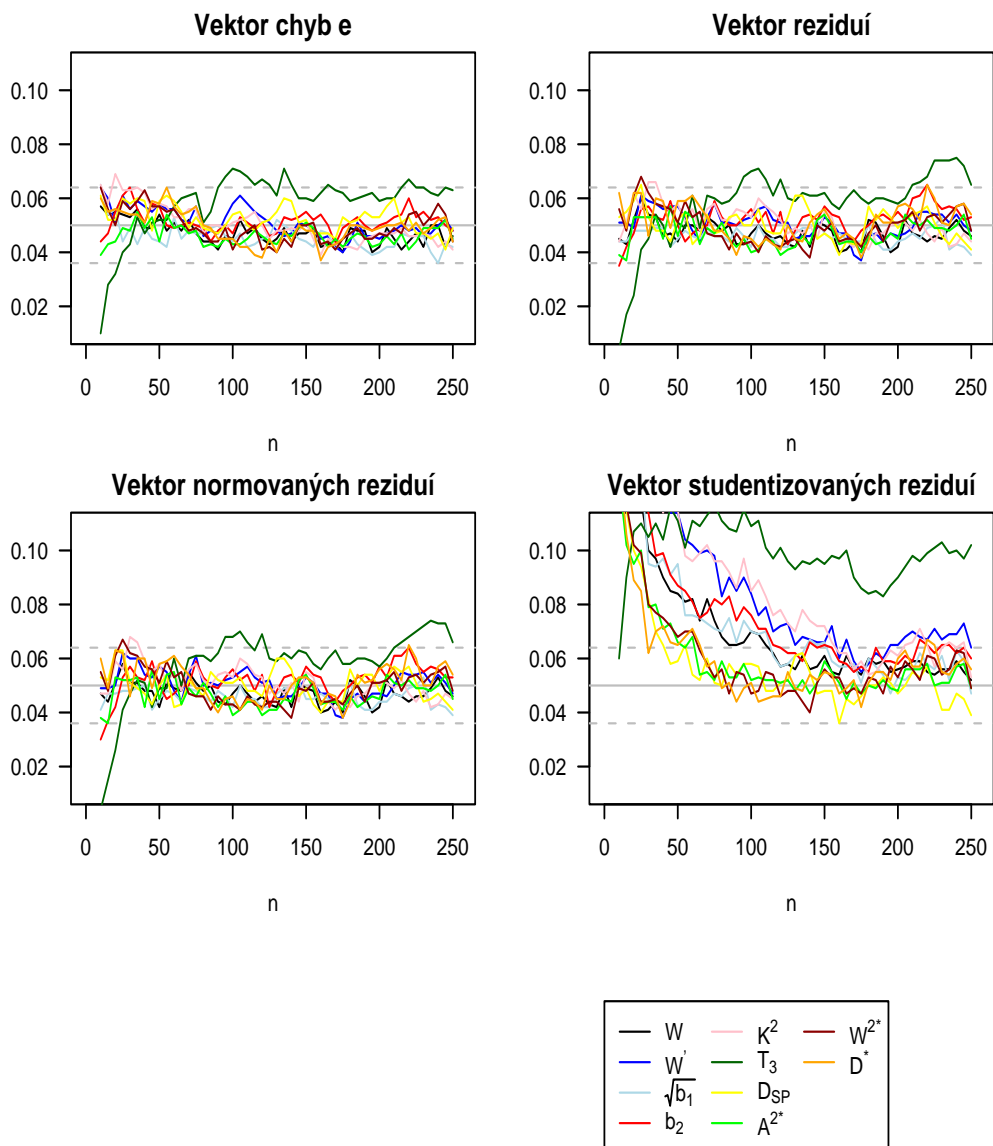
Obrázek A.7: Hladina testů aplikovaných na jednotlivá rezidua v modelu regresní přímky



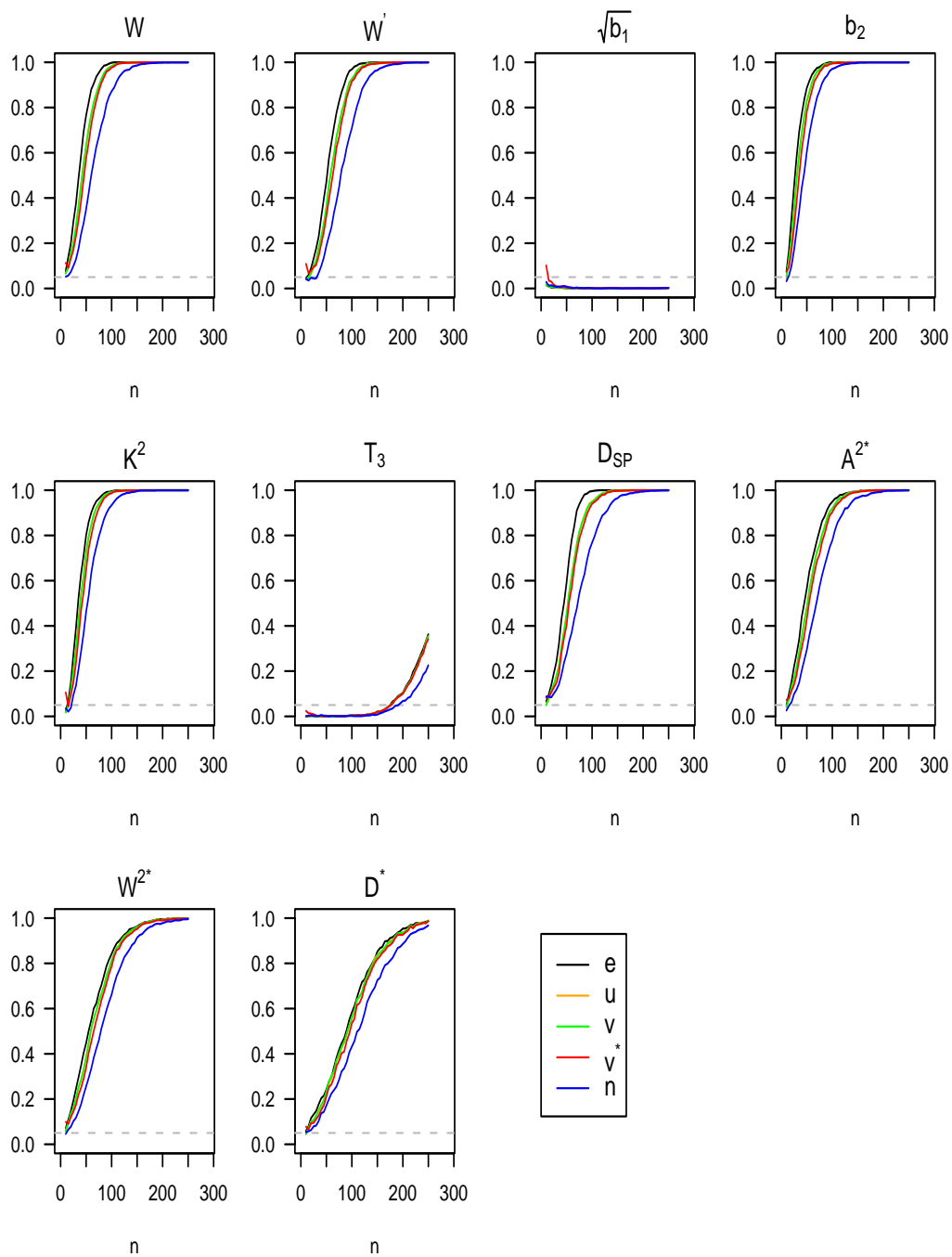
Obrázek A.8: Hladina testů aplikovaných na jednotlivá rezidua v modelu kvadratické závislosti



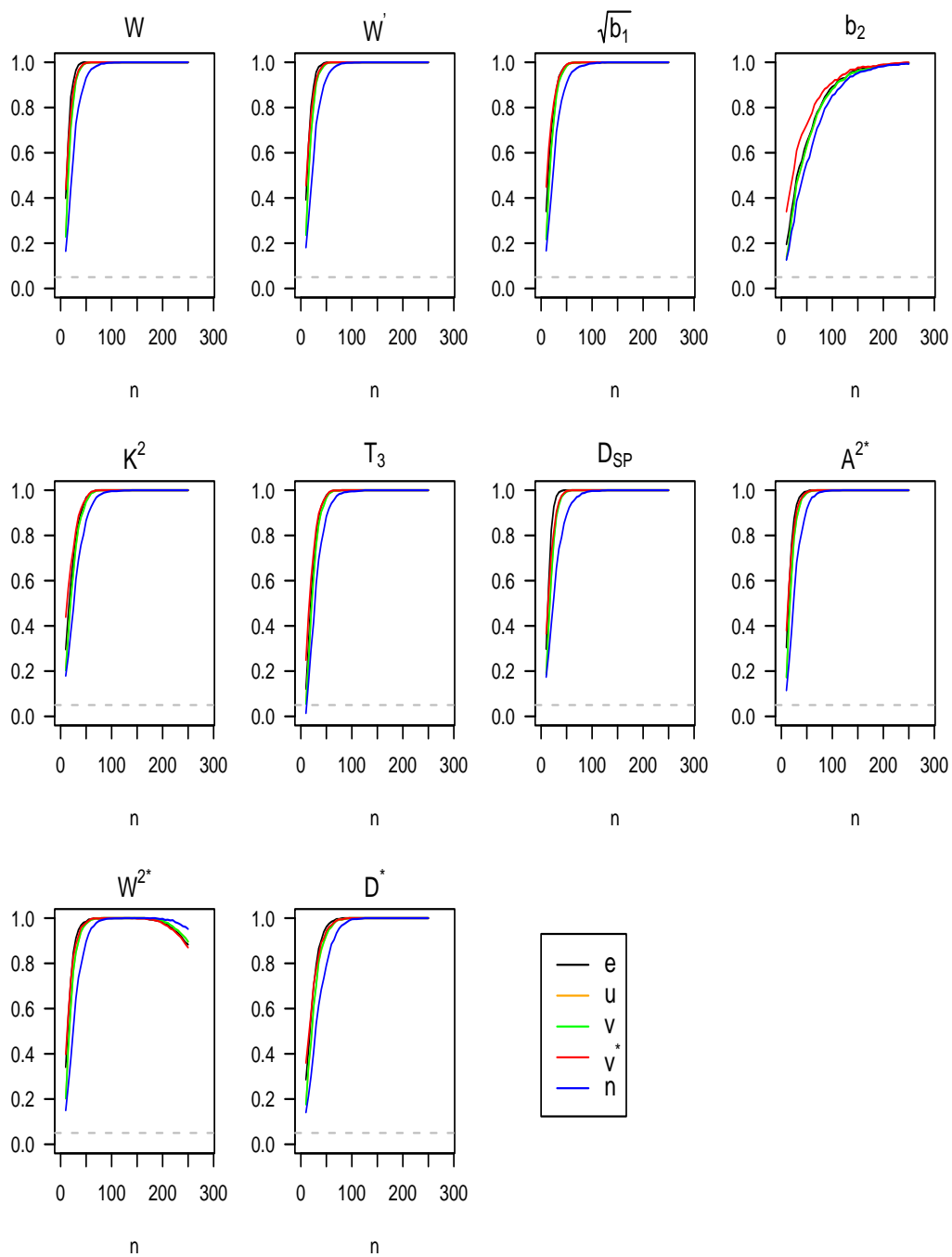
Obrázek A.9: Hladina testů aplikovaných na jednotlivá rezidua v modelu lineární regrese se třemi regresory



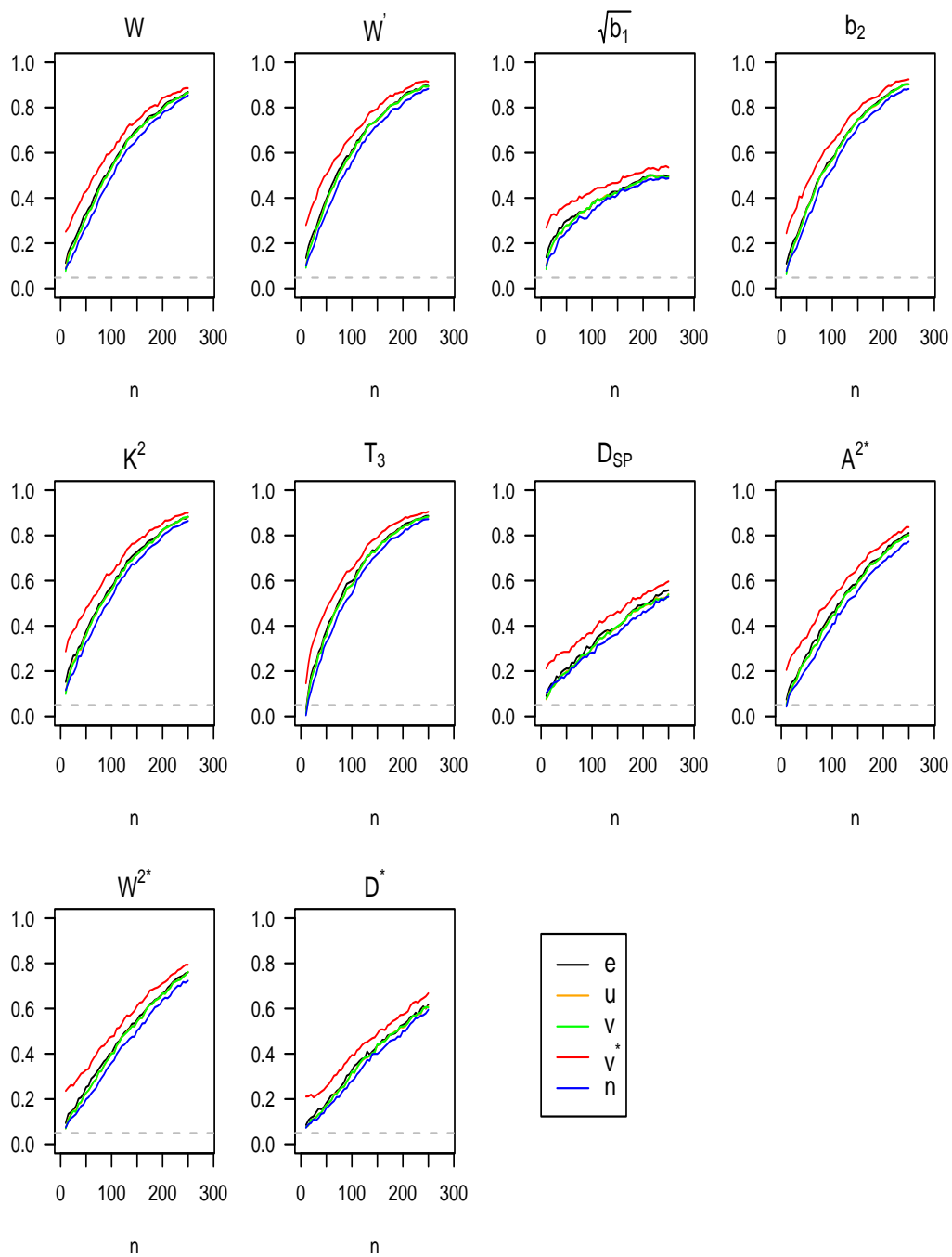
Obrázek A.10: Hladina testů aplikovaných na jednotlivá rezidua v modelu jednoduchého třídění



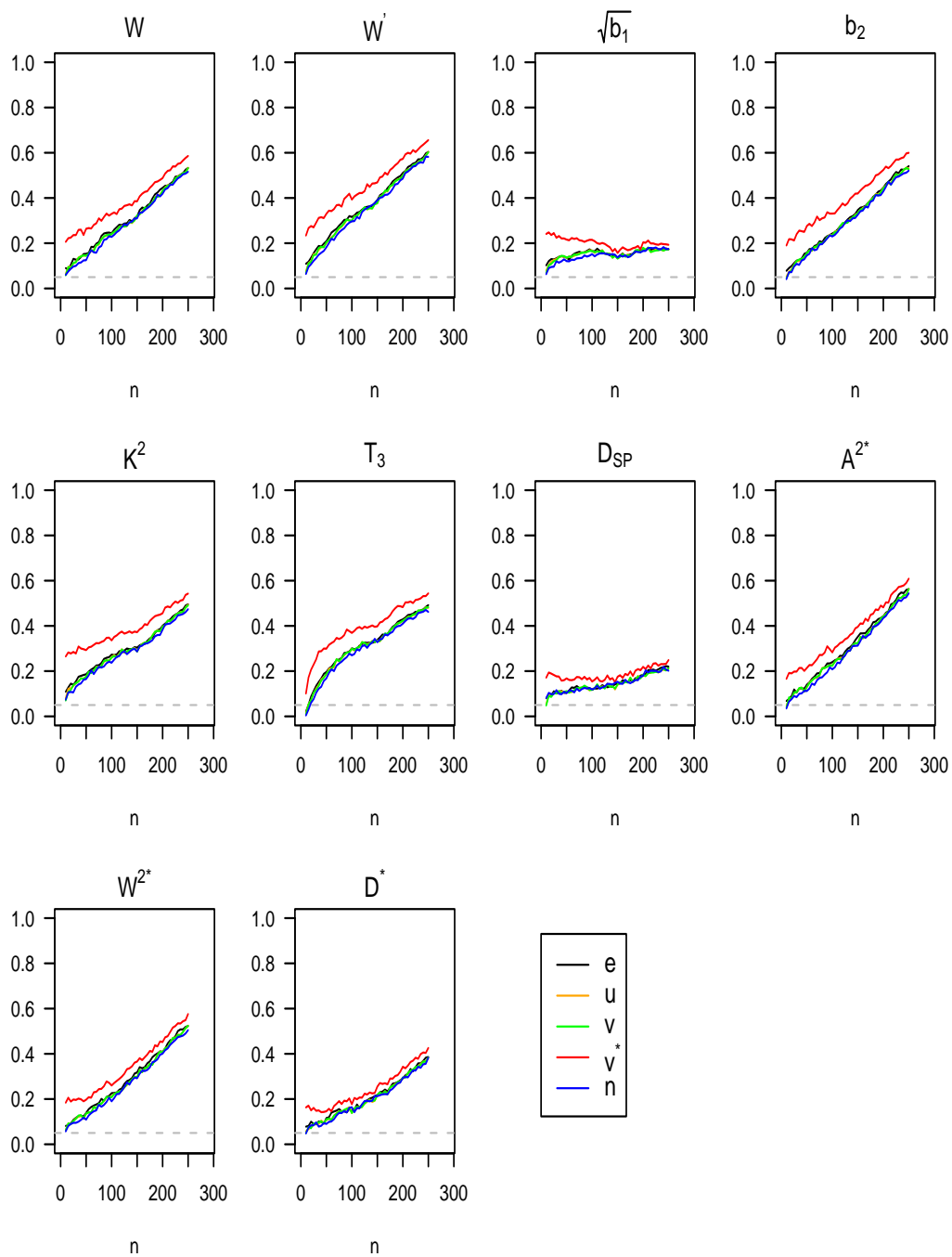
Obrázek A.11: Síla testů aplikovaných na jednotlivá rezidua v modelu regrese pro rozdělení $R(-\sqrt{3}, \sqrt{3})$



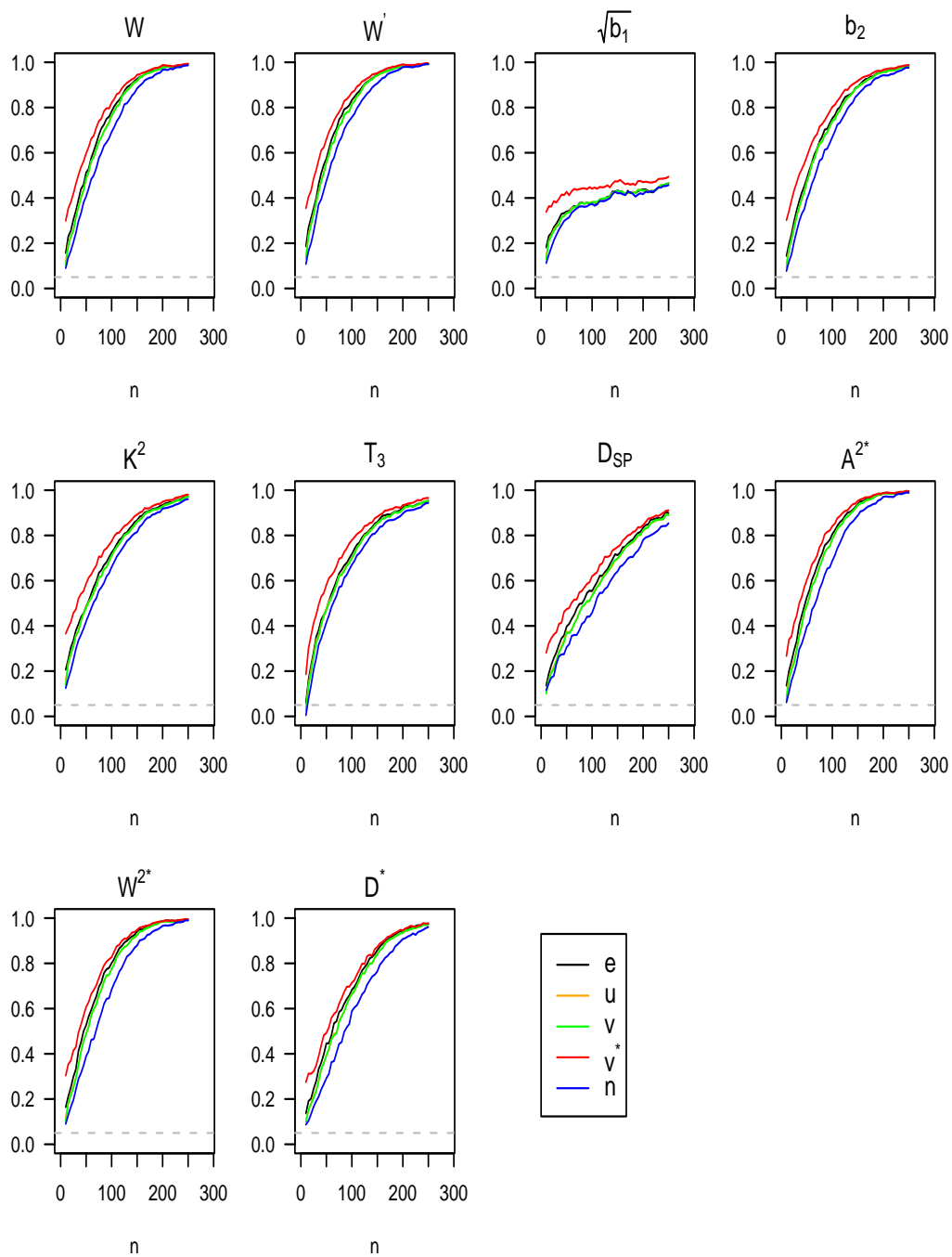
Obrázek A.12: Síla testů aplikovaných na jednotlivá rezidua v modelu regresní přímky pro rozdělení $Exp(2)$



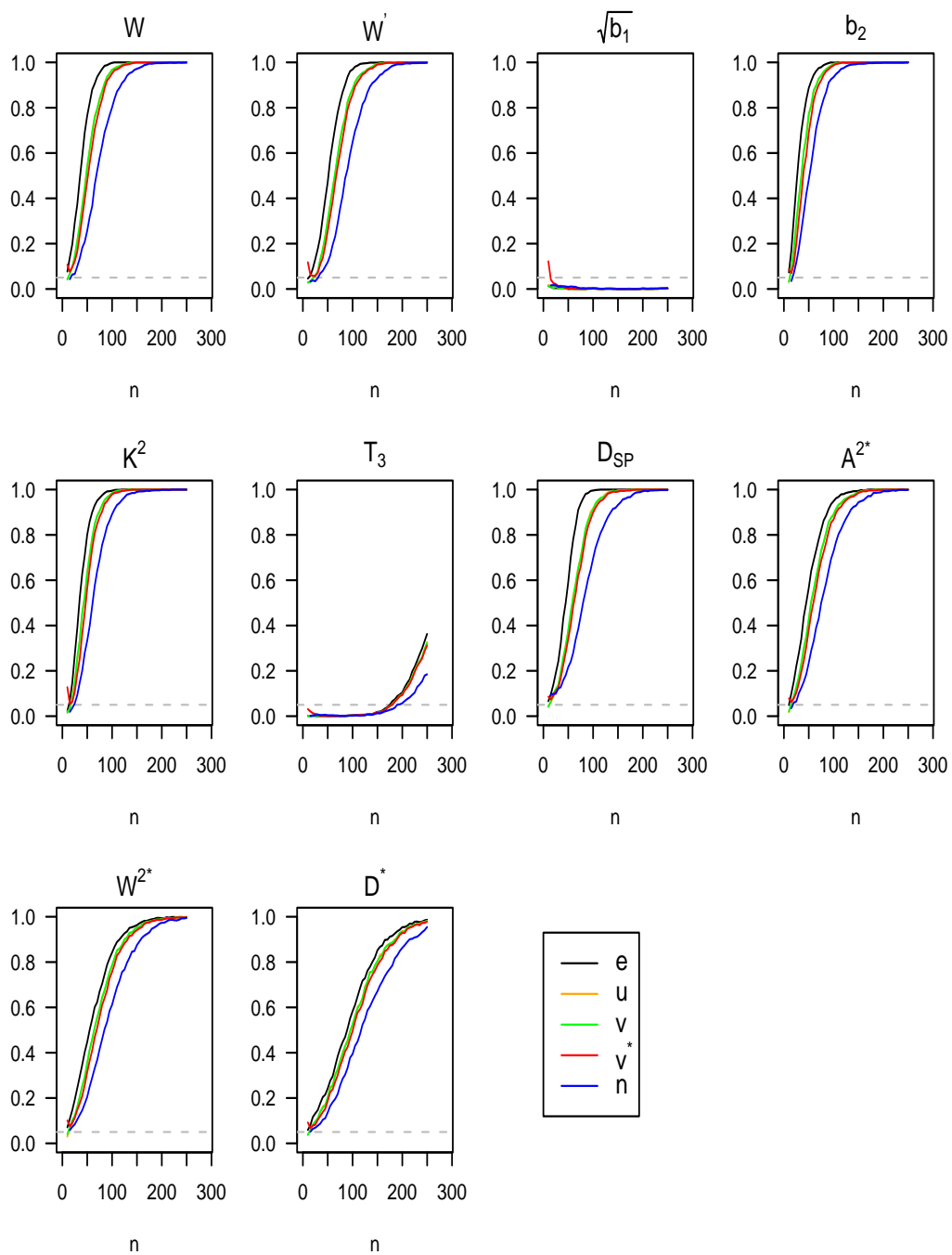
Obrázek A.13: Síla testů aplikovaných na jednotlivá rezidua v modelu regresní přímky pro rozdělení t_5



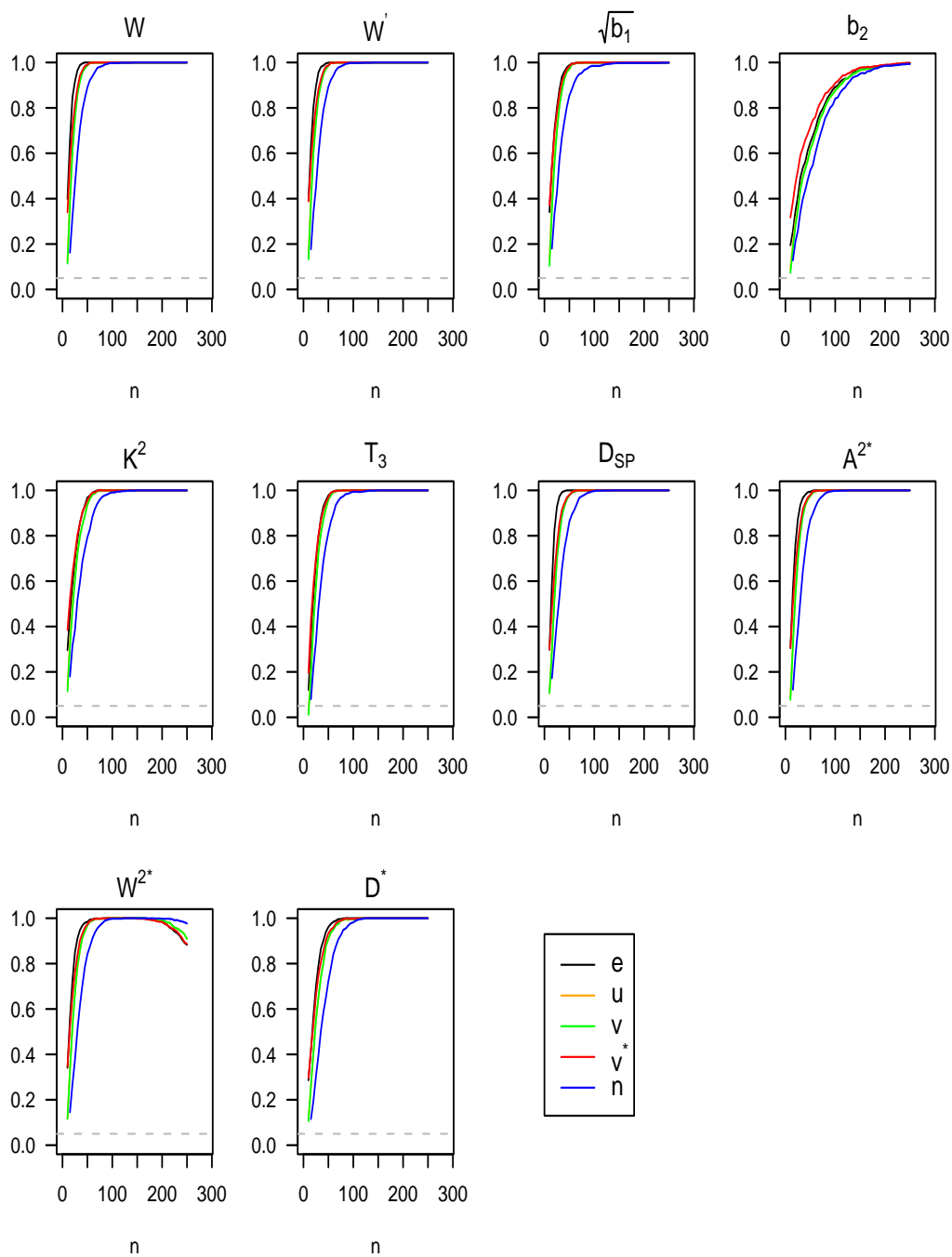
Obrázek A.14: Síla testů aplikovaných na jednotlivá rezidua v modelu regresi přímky pro smíšené normální rozdělení



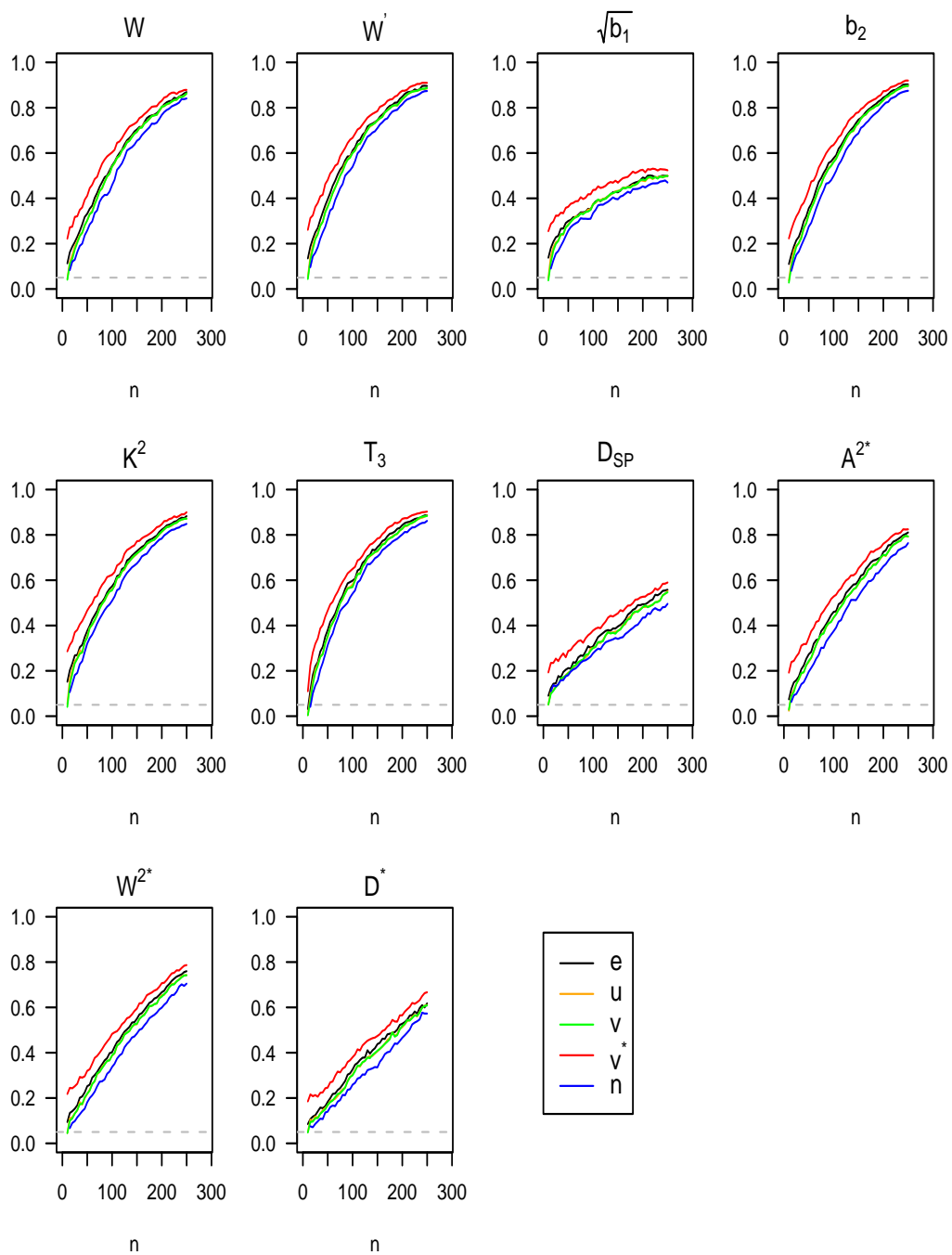
Obrázek A.15: Síla testů aplikovaných na jednotlivá rezidua v modelu regresi přímky pro rozdělení $DExp(1, \frac{1}{\sqrt{2}})$



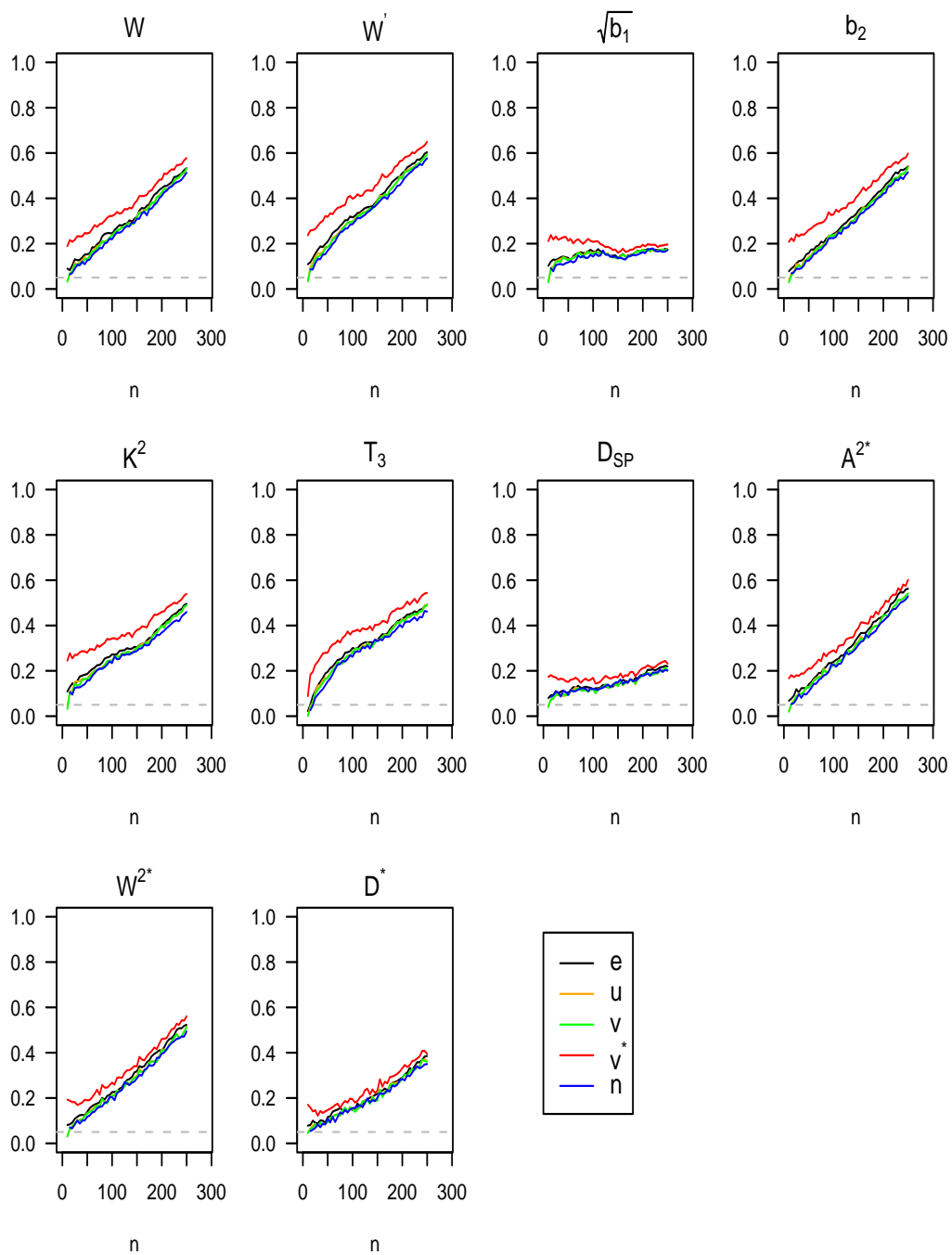
Obrázek A.16: Síla testů aplikovaných na jednotlivá rezidua v modelu kvadratické závislosti pro rozdělení $R(-\sqrt{3}, \sqrt{3})$



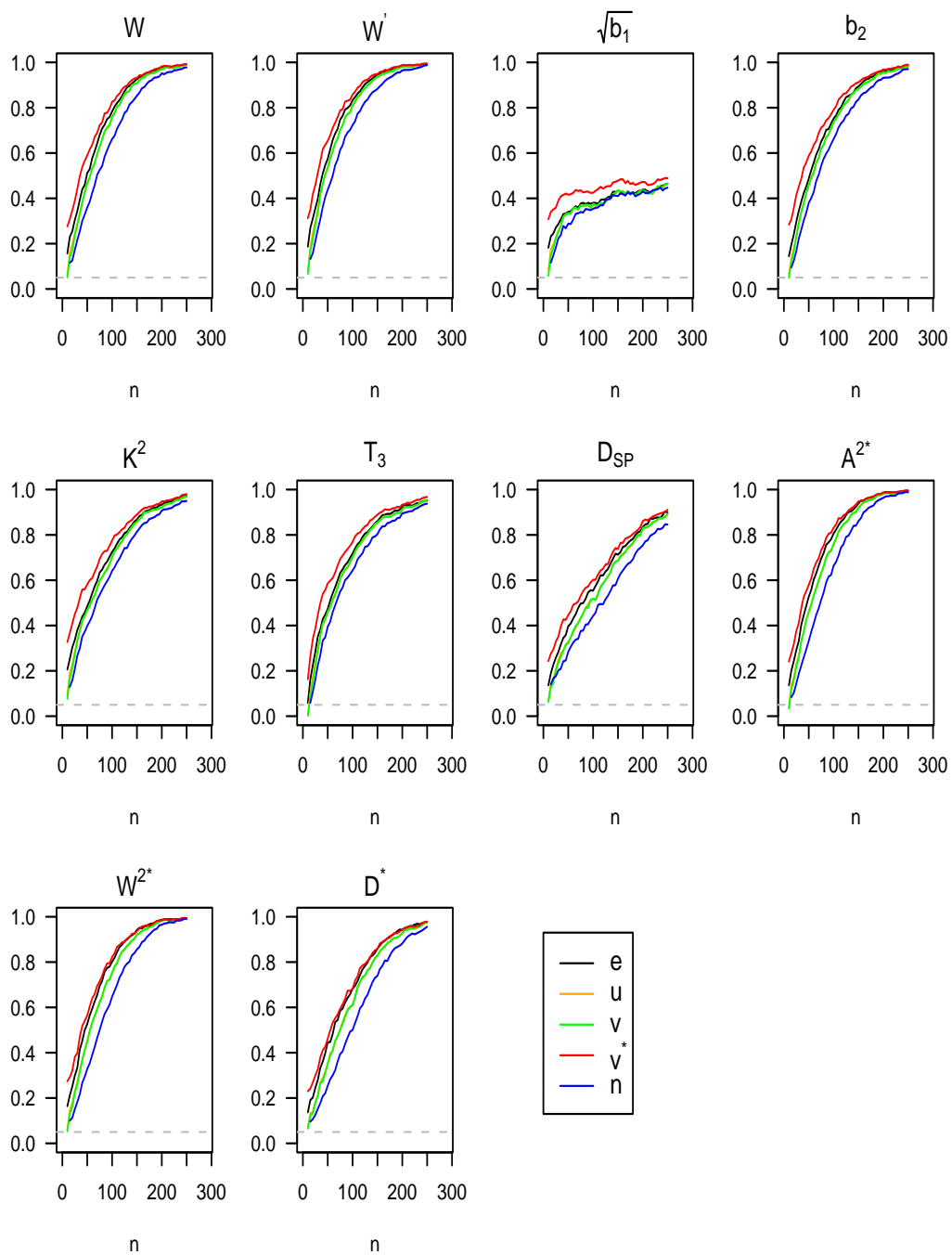
Obrázek A.17: Síla testů aplikovaných na jednotlivá rezidua v modelu kvadratické závislosti pro rozdělení $Exp(2)$



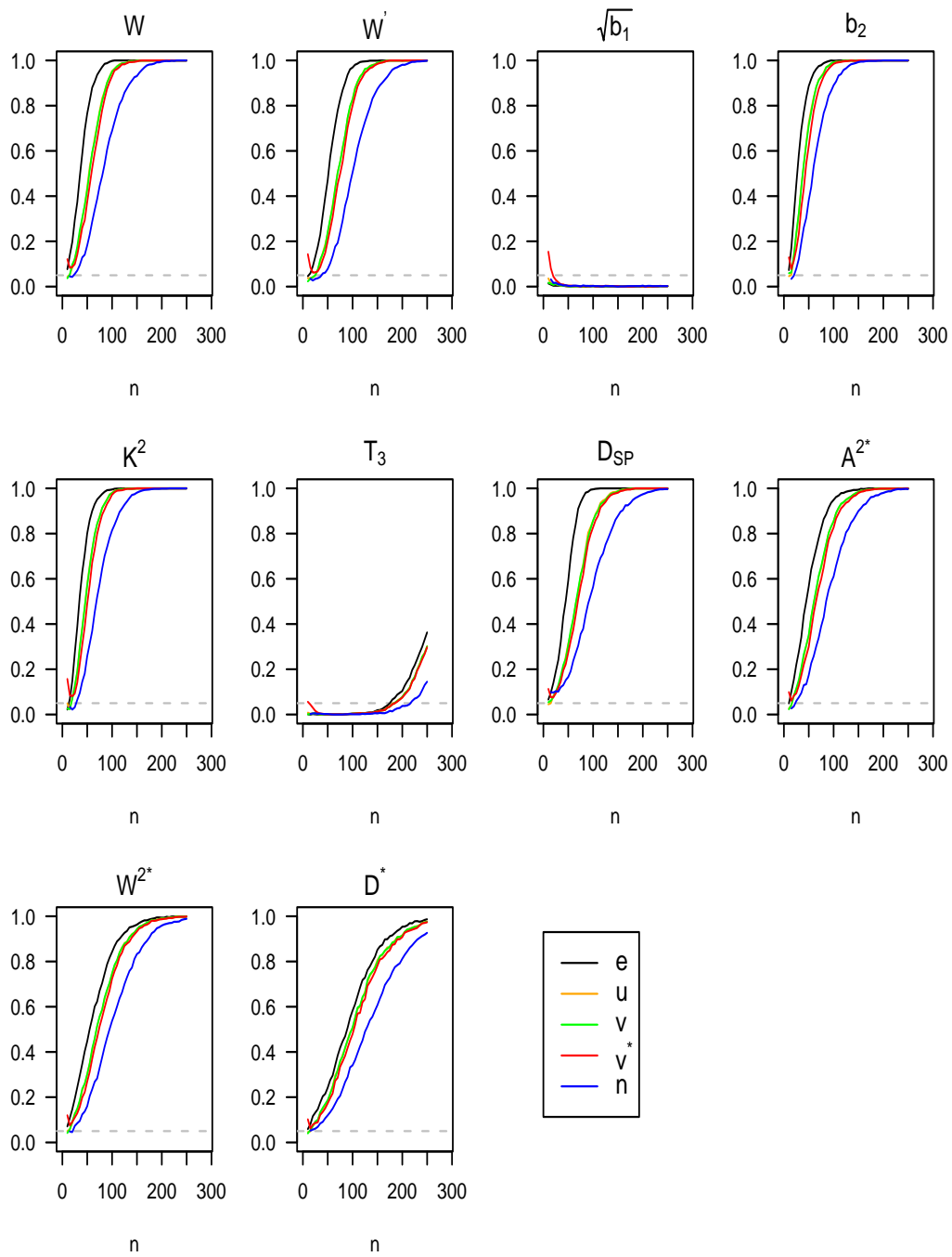
Obrázek A.18: Síla testů aplikovaných na jednotlivá rezidua v modelu kvadratické závislosti pro rozdělení t_5



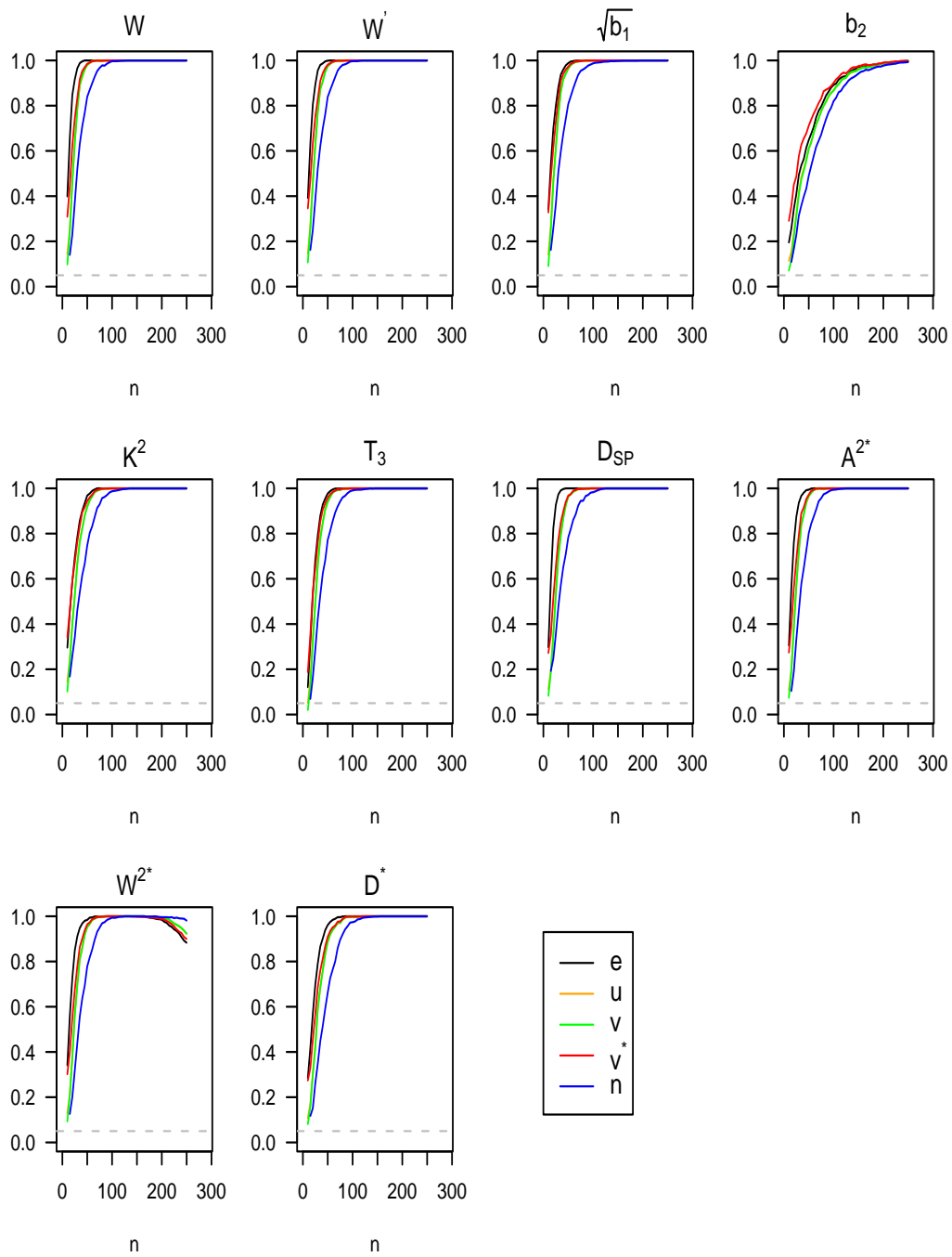
Obrázek A.19: Síla testů aplikovaných na jednotlivá rezidua v modelu kvadratické závislosti pro směs normálních rozdělení



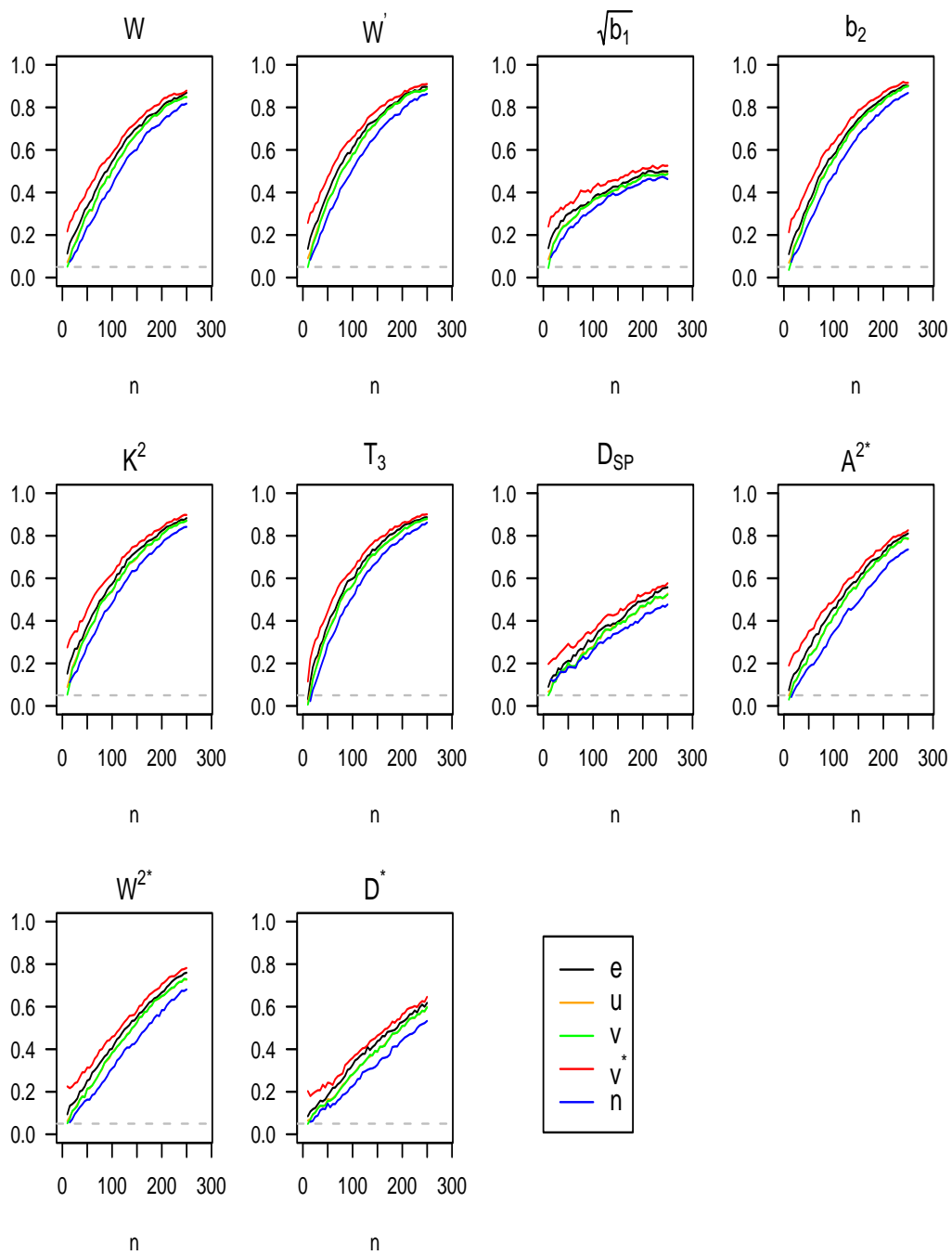
Obrázek A.20: Síla testů aplikovaných na jednotlivá rezidua v modelu kvadratické závislosti pro rozdělení $DExp(1, \frac{1}{\sqrt{2}})$



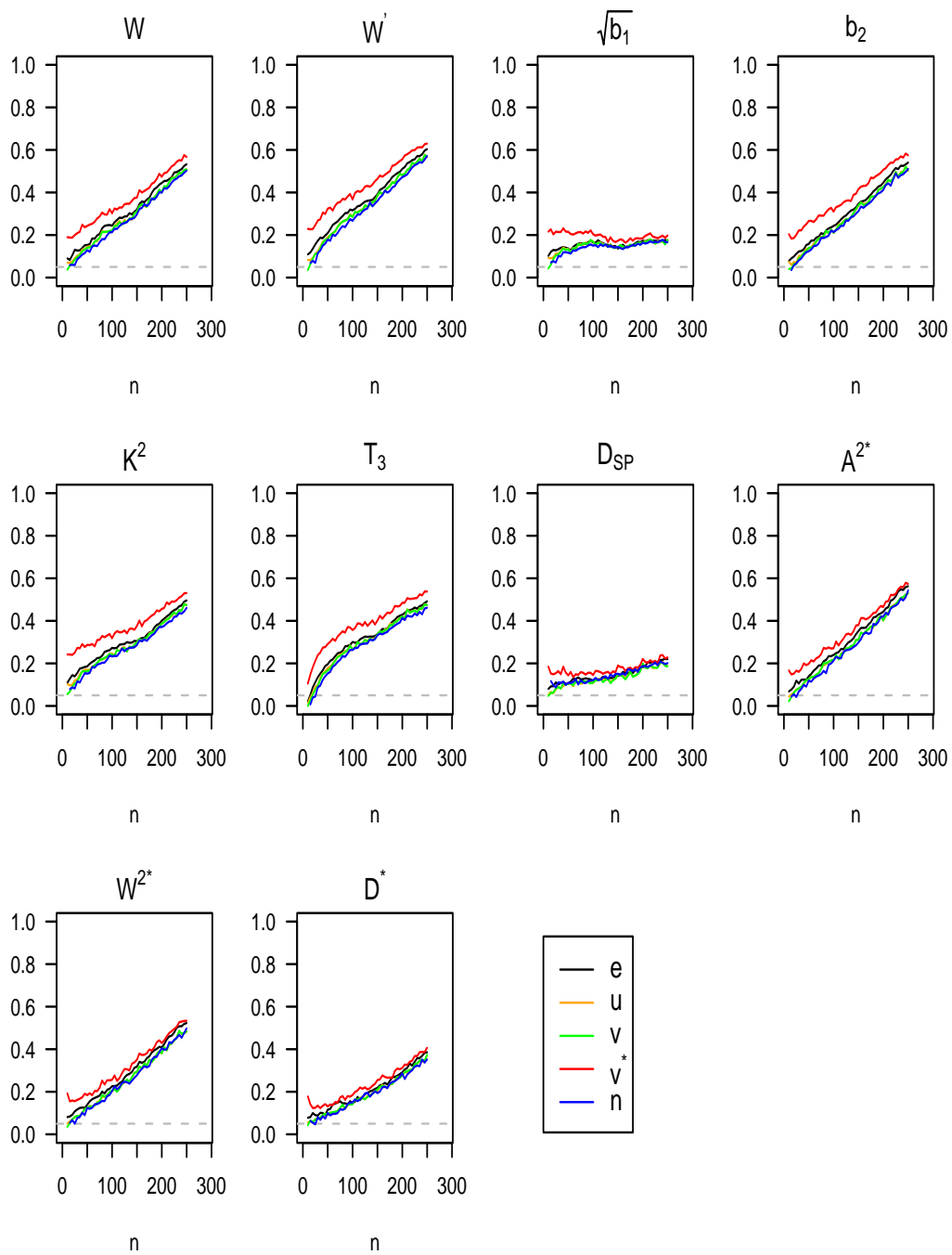
Obrázek A.21: Síla testů aplikovaných na jednotlivá rezidua v modelu lineární regrese se třemi regresory pro rozdělení $R(-\sqrt{3}, \sqrt{3})$



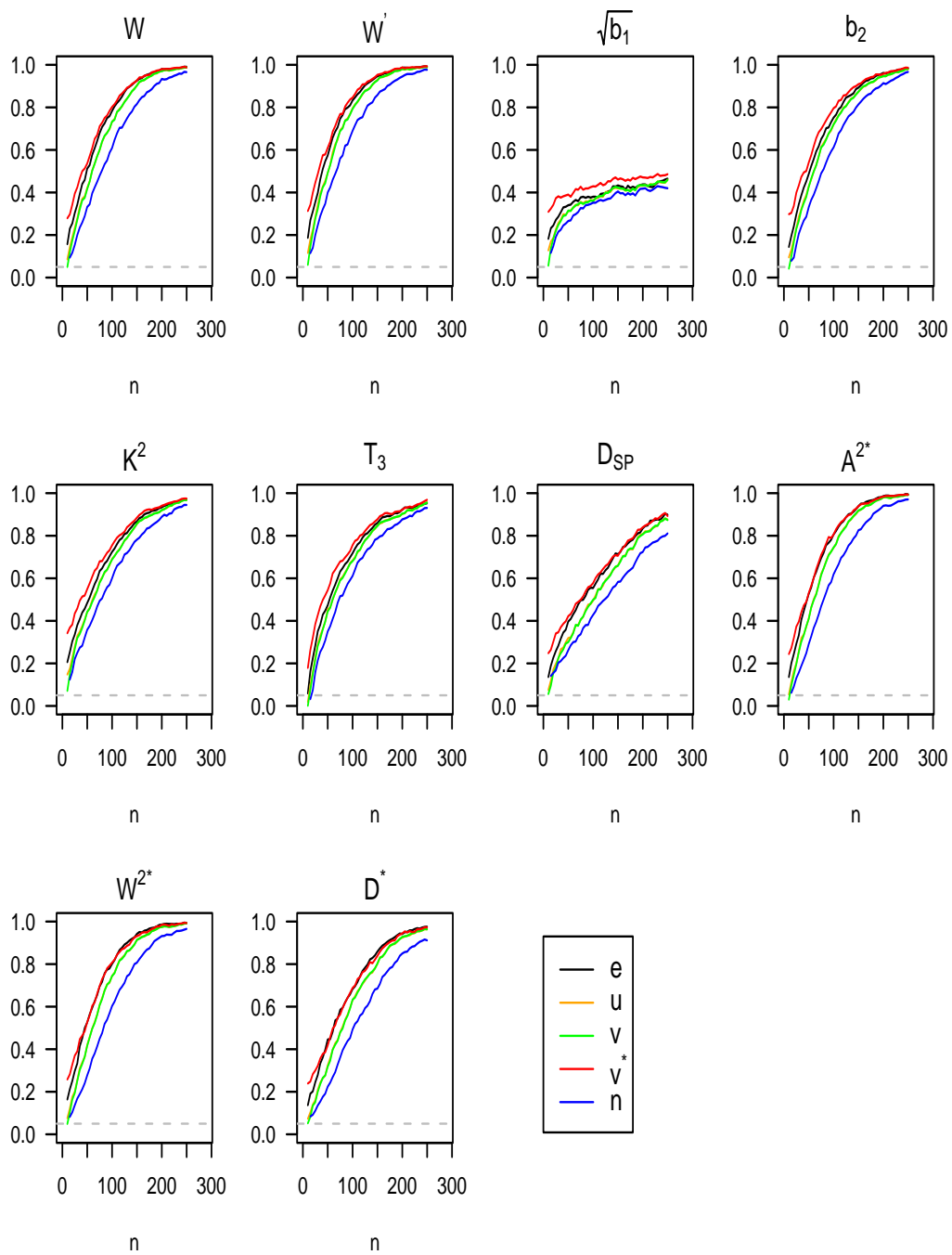
Obrázek A.22: Síla testů aplikovaných na jednotlivá rezidua v modelu lineární regrese se třemi regresory pro rozdělení $Exp(2)$



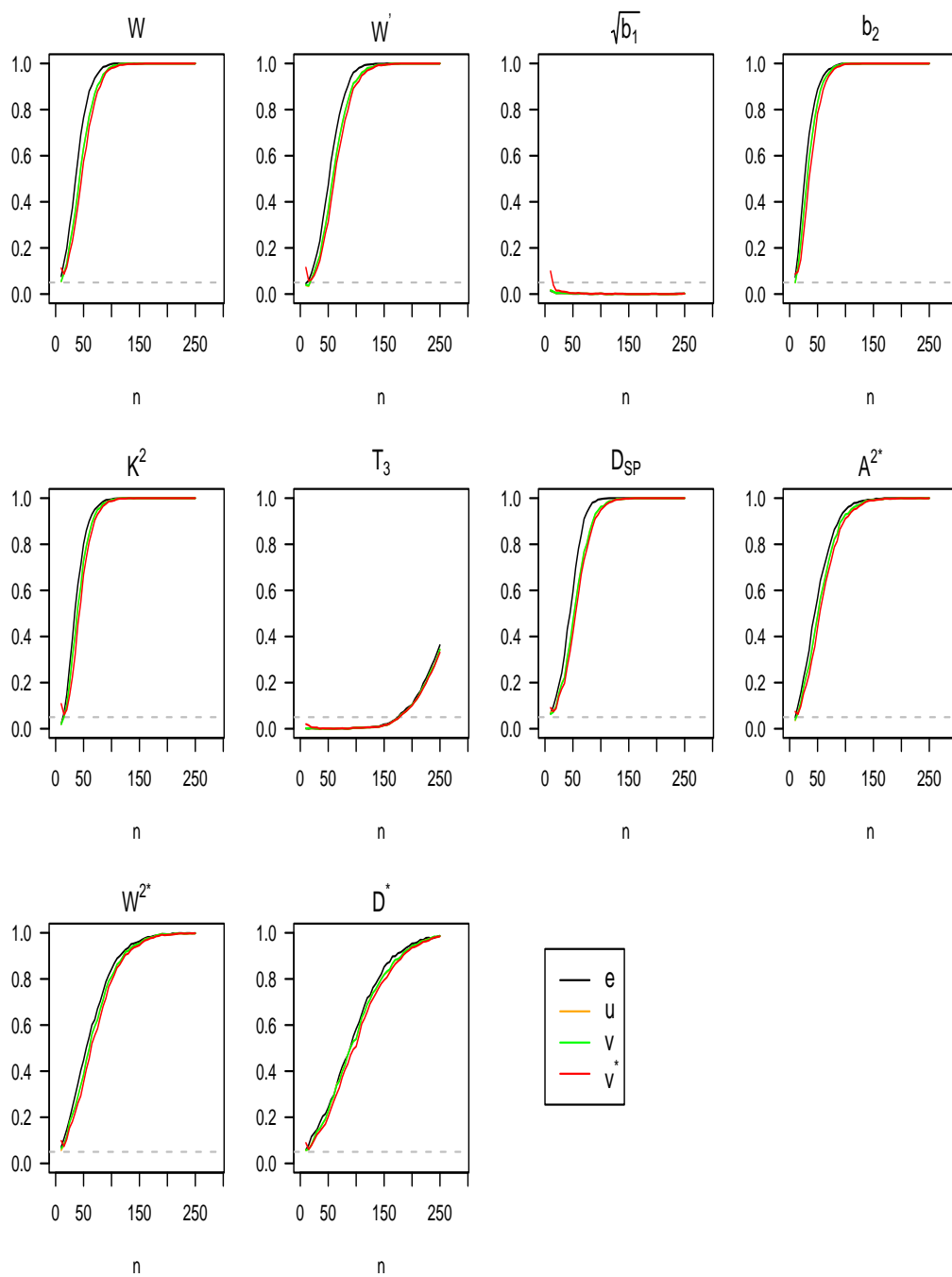
Obrázek A.23: Síla testů aplikovaných na jednotlivá rezidua v modelu lineární regrese se třemi regresory pro rozdělení t_5



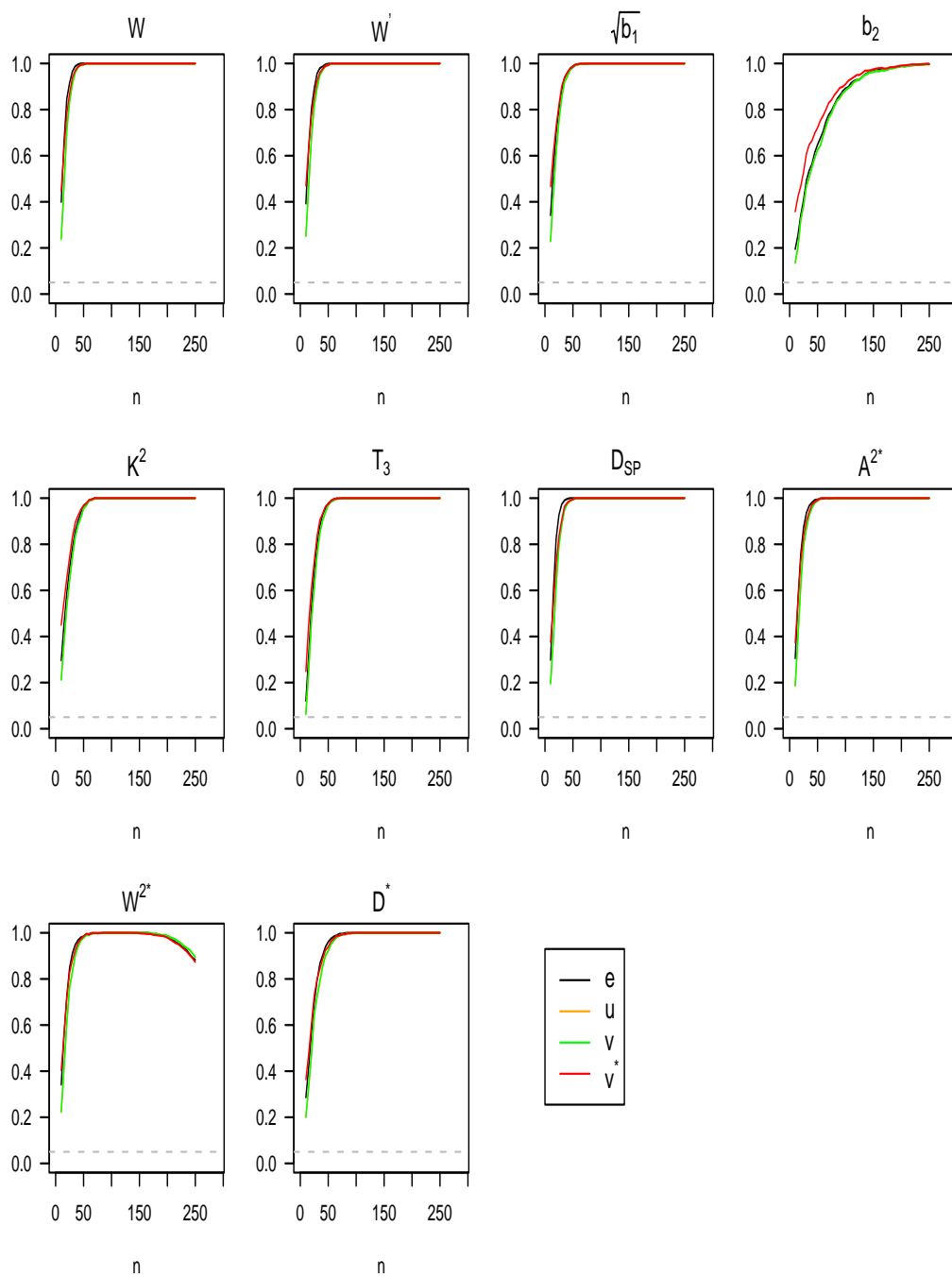
Obrázek A.24: Síla testů aplikovaných na jednotlivá rezidua v modelu lineární regrese se třemi regresory pro smíšené normální rozdělení



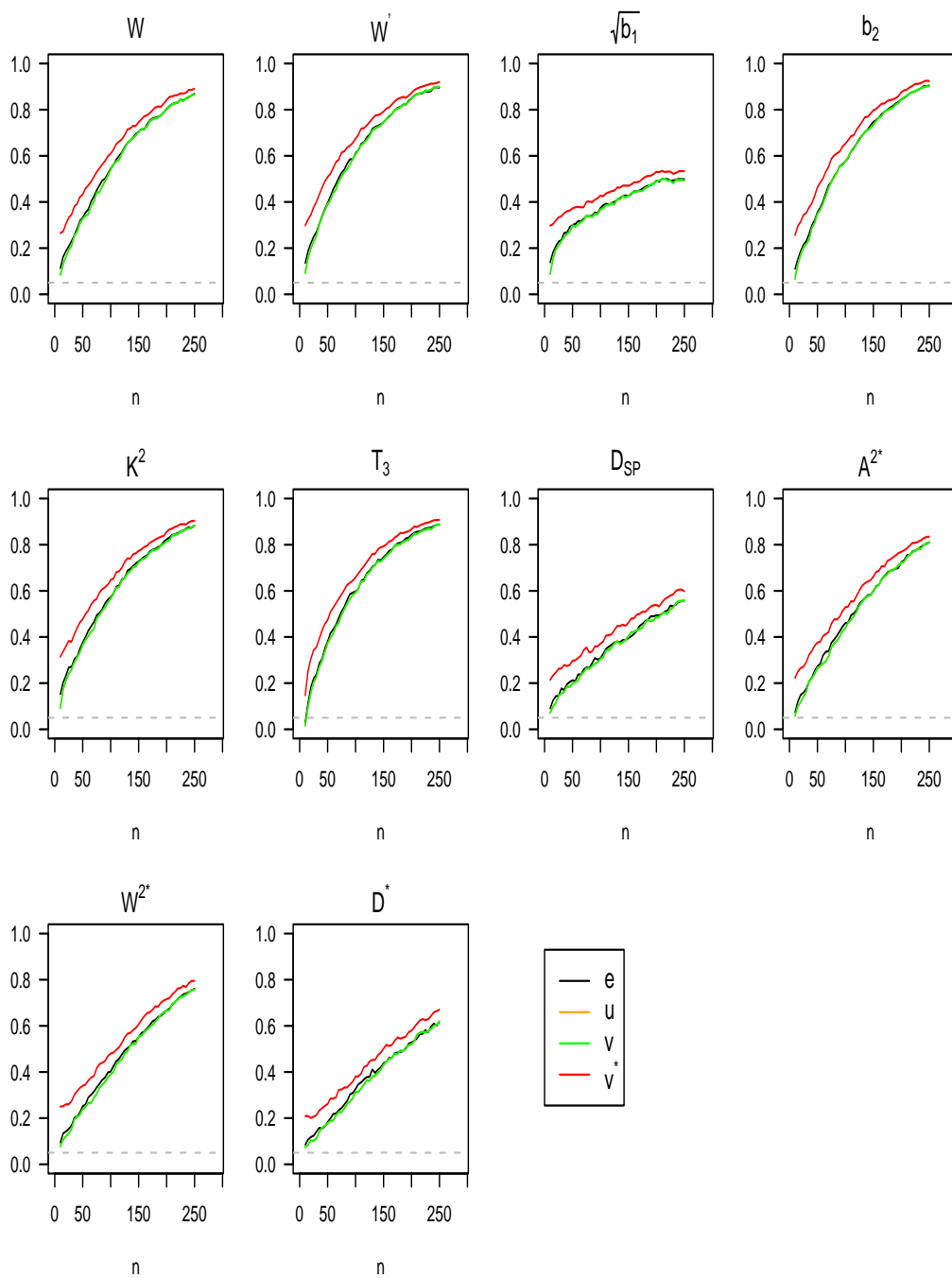
Obrázek A.25: Síla testů aplikovaných na jednotlivá rezidua v modelu lineární regrese se třemi regresory pro rozdělení $DExp(1, \frac{1}{\sqrt{2}})$



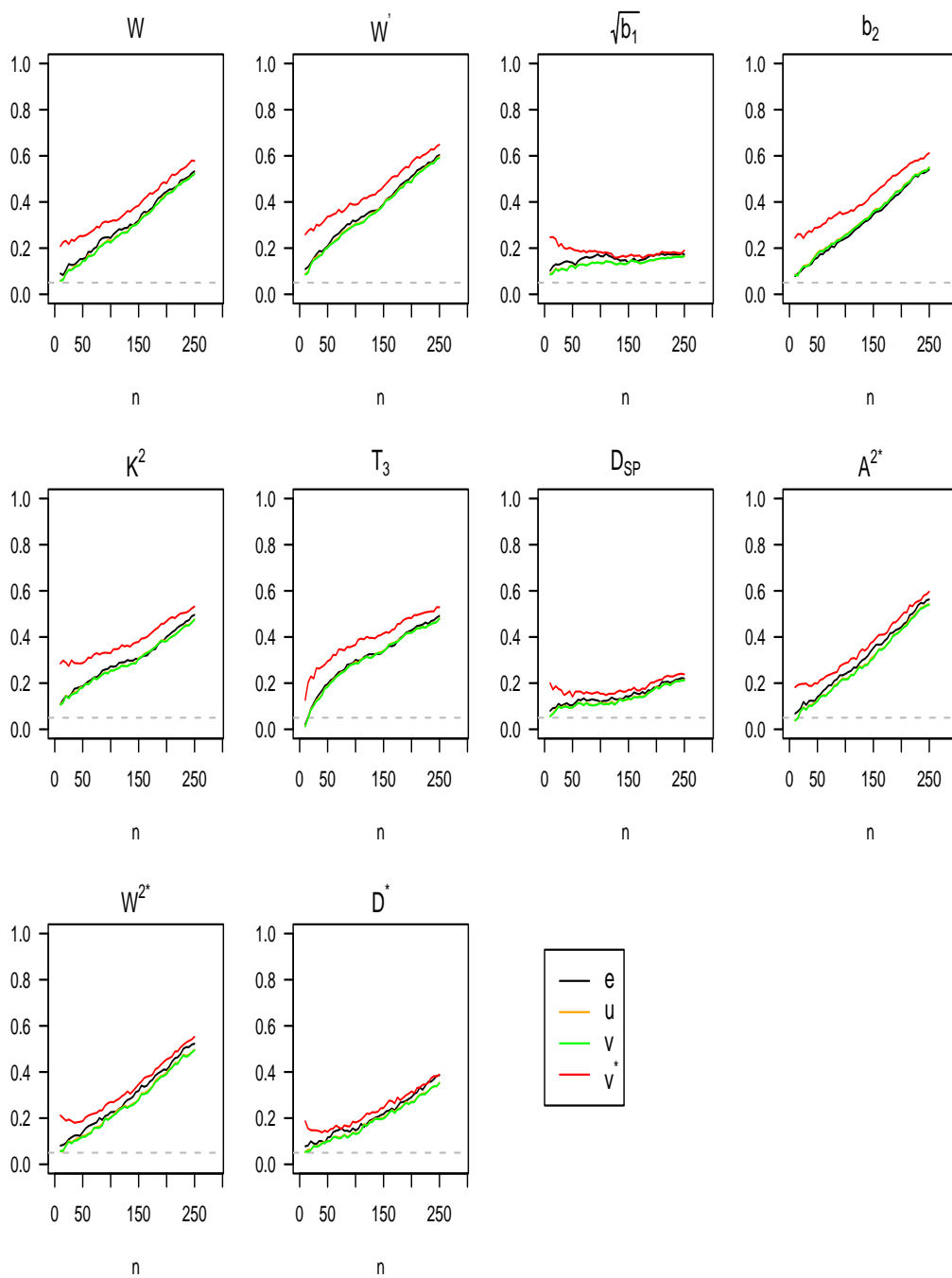
Obrázek A.26: Síla testů aplikovaných na jednotlivá rezidua v modelu jednoduchého třídění pro rozdělení $R(-\sqrt{3}, \sqrt{3})$



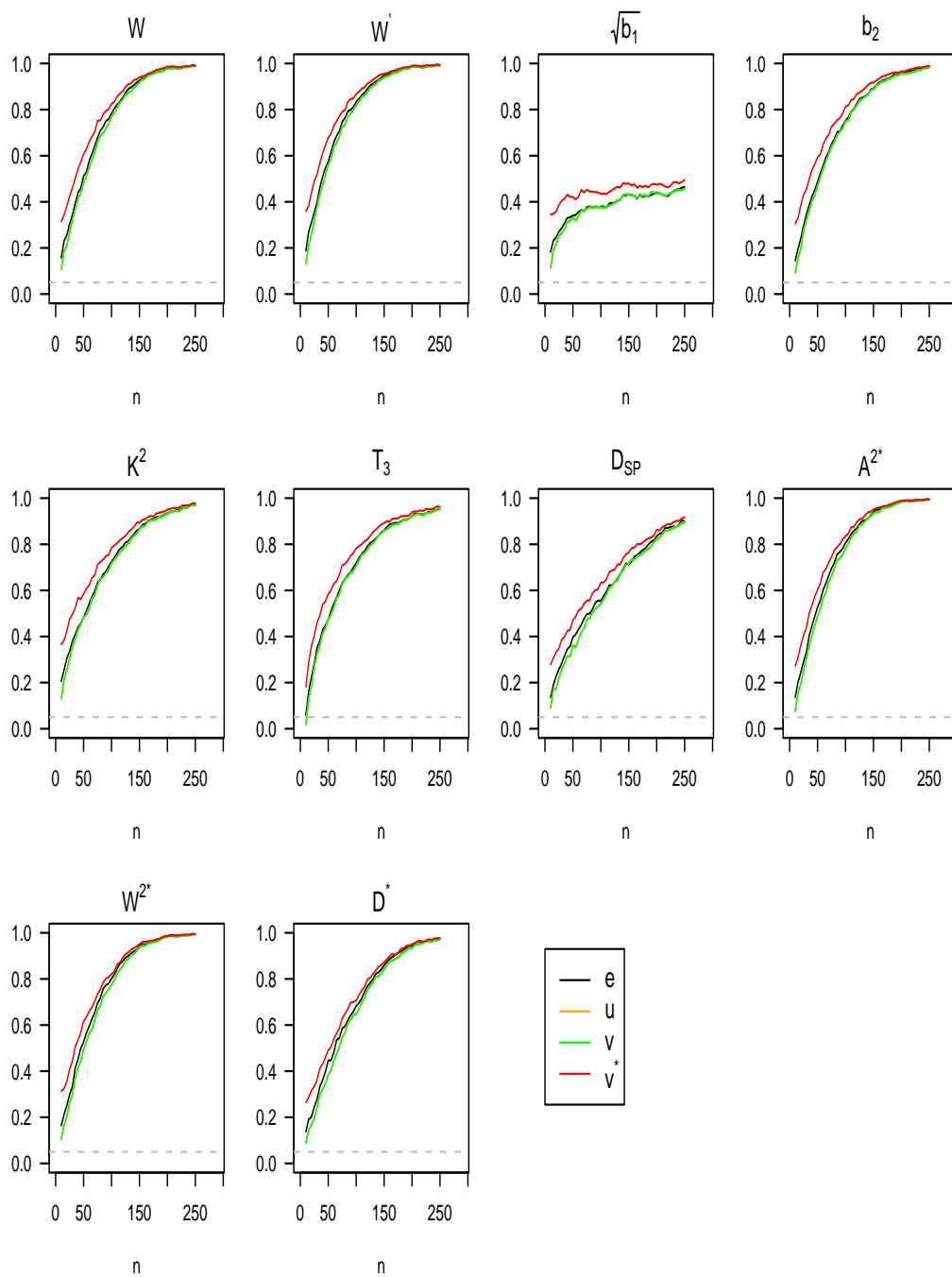
Obrázek A.27: Síla testů aplikovaných na jednotlivá rezidua v modelu jednoduchého třídění pro rozdělení $Exp(2)$



Obrázek A.28: Síla testů aplikovaných na jednotlivá rezidua v modelu jednoduchého třídění pro rozdělení t_5



Obrázek A.29: Síla testů aplikovaných na jednotlivá rezidua v modelu jednoduchého třídění pro smíšené normální rozdělení



Obrázek A.30: Síla testů aplikovaných na jednotlivá rezidua v modelu jednoduchého třídění pro rozdělení $DExp(1, \frac{1}{\sqrt{2}})$

Reference

- [1] Jiří Anděl. *Základy matematické statistiky*. Preprint, UK MFF, Praha, 2002.
- [2] Jaromír Antoch, Dana Vorlíčková. *Vybrané metody statistické analýzy dat*. Academia Praha, 1992.
- [3] Ralph B. D'Agostino, Albert Belanger, and Ralph B. Jr. D'Agostino. A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4):316–321, November 1990.
- [4] Sucharita Ghosh. A new graphical tool to detect non-normality. *Journal of the Royal Statistical Society. Series B*, 58(4):691–702, 1996.
- [5] John R. Michael. The stabilized probability plot. *Biometrika*, 70(1):11–17, April 1983.
- [6] Patrick Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31(2):115–124, 1982.
- [7] Patrick Royston. Graphical detection of non-normality by using Michael's statistic. *Applied Statistic*, 42(1):153–158, 1993.
- [8] Patrick Royston. Remark AS R94: A remark on algorithm AS 181: The W test for normality. *Applied Statistics*, 44(4):547–551, 1995.
- [9] S. S. Shapiro and S. R. Francia, An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, March 1972.
- [10] S. S. Shapiro and M. B. Wilk, An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, December 1965.
- [11] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, September 1974.
- [12] M. Štefek. *Porušení předpokladu o normálním rozdělení v lineárním modelu*. Diplomová práce, UK MFF, Praha, 1994.
- [13] Henry C. Thode. *Testing for normality*. Marcel Dekker, New York, 2002.
- [14] Karel Zvára. *Učební text k regresi*. 2006.