Metoda backward výběru proměnných v lineární regresi a její vlastnosti

Aktuárský seminář, 13. dubna 2018

Milan Bašta

1 Metody výběru proměnných do modelu

2 Monte Carlo simulace, backward metoda

3 Pravděpodobnosti různých finálních modelů

4 Inference

5 Přesnost odhadu regresní funkce

6 Rozšíření výsledků

Postupné odebírání vysvětlujících proměnných

- Angl. **backward elimination**.
- V modelu máme všechny vysvětlující proměnné. Postupně odebíráme takové vysvětlující proměnné, jejichž p-hodnota (příslušející dílčímu t-testu o nulové hodnotě parametru v aktuálním modelu) je největší ze všech vysvětlujících proměnných v modelu a zároveň větší než zvolená hladina α_{remove} (např. 0.05, 0.1 apod.). Po odebrání takovéto vysvětlující proměnné z modelu se p-hodnoty u ostatních vysvětlujících proměnných, které v modelu zbyly, v obecnosti změní.
- Tuto proceduru postupného odebírání opakujeme do té doby, než všechny vysvětlující proměnné, které v modelu zbyly, mají p-hodnotu menší nebo rovnu α_{remove}.

Postupné zařazování vysvětlujících proměnných

- Angl. **forward selection**.
- Začneme s modelem, ve kterém není žádná vysvětlující proměnná. Pro každou vysvětlující proměnnou, která není v modelu, vypočteme p-hodnotu (na základě dílčího t-testu) odpovídající přidání této proměnné do modelu. Ze všech takovýchto proměnných posléze do modelu přidáme tu s nejnižší p-hodnotou, pokud je tato p-hodnota zároveň nižší než zvolená hladina α_{add}.
- Tento postup opakujeme do té doby, dokud existují vysvětlující proměnné, které můžeme přidávat.

Krokové metody v literatuře

- Derksen, S., & Keselman, H. J. (1992): Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45(2), 265-282.
- Harrell, F. (2001): Regression modeling strategies, with applications to linear models, logistic regression and survival analysis. Springer.
- Molodkina, K. (2014): Krokové metody v lineární regresi a jejich vlastnosti. Bakalářská práce MFF UK.
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006): Why do we still use stepwise modelling in ecology and behaviour? Journal of animal ecology, 75(5), 1182-1189.

Některé alternativy ke krokovým metodám

- Metoda **best subset**: Procházíme, všechny možné modely a pro každý z modelů vyhodnocujeme kritérium kvality modelu (např. upravený index determinace, AIC, apod.). Následně jako "nejlepší" model vybíráme ten, jež poskytuje optimální hodnotu daného kritéria (největší upravený index determinace, nejnižší hodnotu AIC, apod.).
- **lasso regrese** (Tibshirani, 1996)[1].

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

1 Metody výběru proměnných do modelu

2 Monte Carlo simulace, backward metoda

3 Pravděpodobnosti různých finálních modelů

4 Inference

5 Přesnost odhadu regresní funkce

6 Rozšíření výsledků

Monte Carlo simulace

- Lineární regresní model (bez absolutního členu)

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_m x_{im} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde n je počet pozorování, Y_i je vysvětlovaná proměnná pro i-té pozorování, x_{i1}, x_{i2}, ..., x_{ik} jsou známé konstanty, ε_i je chyba pro i-té pozorování a $1 \leq m \leq k$ je počet regresních parametrů, jež jsou nenulové (viz dále).

- V rámci simulace má vektor chyb

$$[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T$$

n-rozměrné normální rozdělení s nulovým vektorem středních hodnot a kovarianční maticí $\sigma^2 \mathbf{I}$, kde $\sigma^2 = 6.25$.

- V rámci simulace volíme postupně $n = 50, 300$.

Vysvětlující proměnné

- V rámci simulace volíme pevně $k = 20$.
- Empirický průměr hodnot libovolné vysvětlující proměnné je 0.
- Empirický rozptyl hodnot libovolné vysvětlující proměnné (tj. hodnot $x_{1j}, x_{2j}, ..., x_{nj}$, pro $j = 1, ..., k$) je roven 1.
- Histogram hodnot vysvětlujících proměnných je blízký "gaussovskému histogramu".
- Empirická korelace mezi hodnotami libovolné dvojice různých vysvětlujících proměnných je rovna ρ. V simulaci volíme postupně $\rho = 0, 0.8$.

Volba regresních parametrů

- Volíme postupně $m = 5, 15$.
- Regresní parametry jsou dány jako:

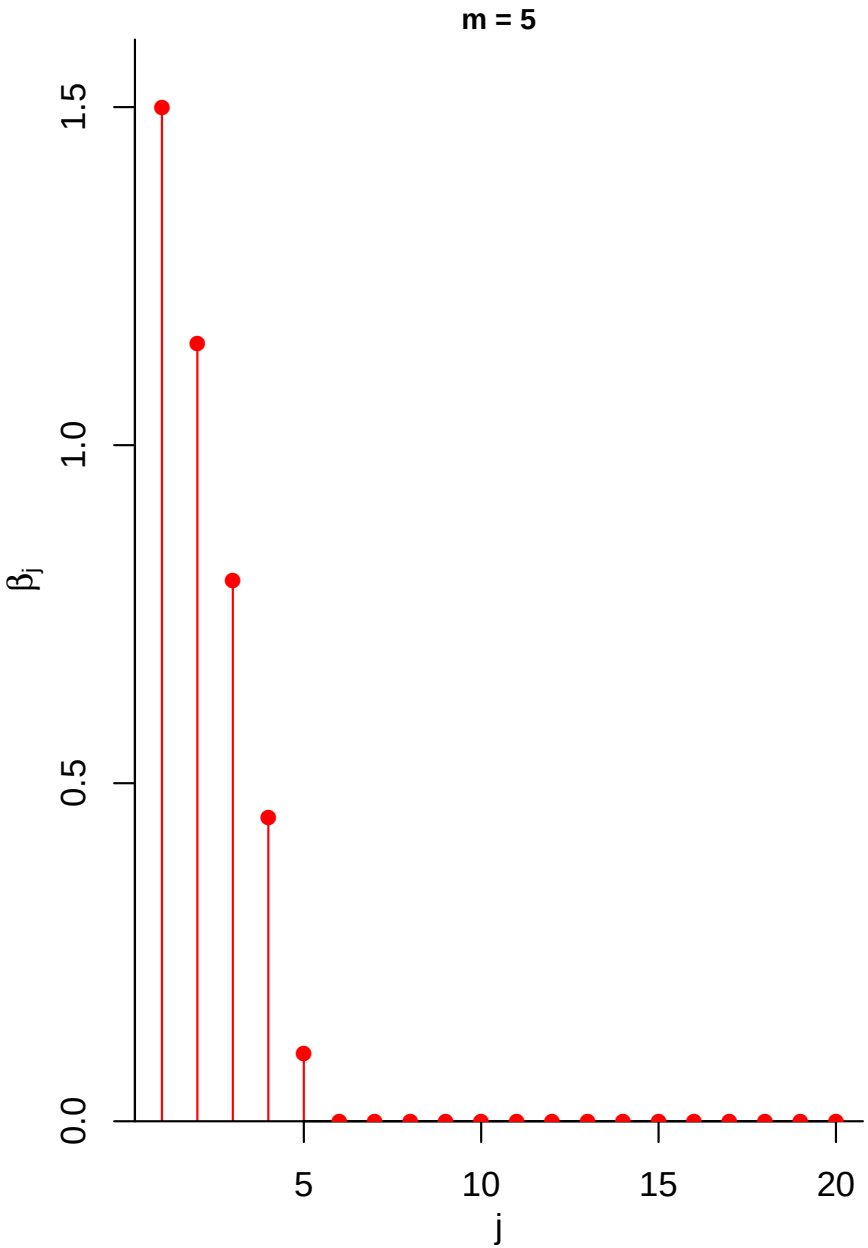
$$\beta_j = \begin{cases} 1.5 - 1.4\frac{j-1}{m-1}, & \text{pro } j = 1, 2, ..., m, \\ 0, & \text{pro } j = m+1, ..., k. \end{cases}$$

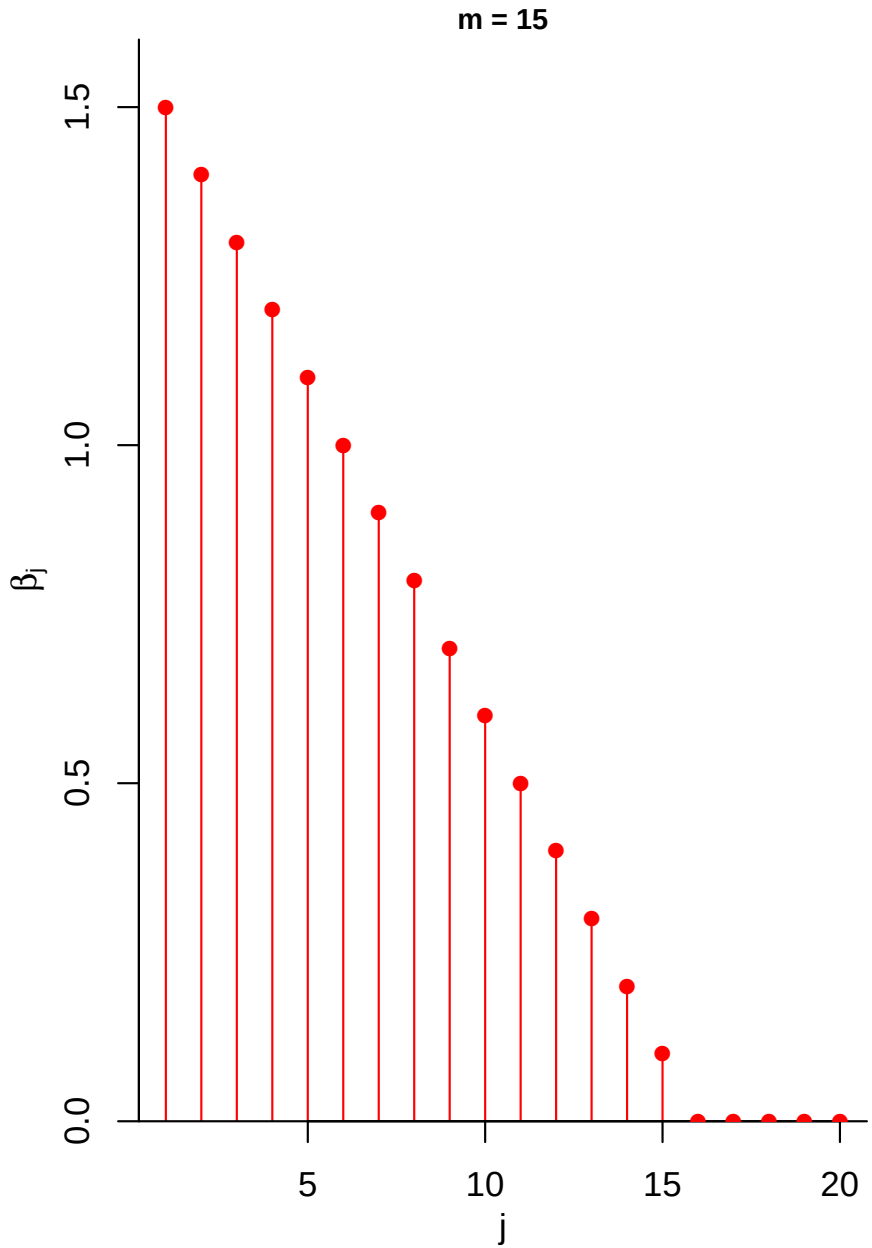
- Pro $m = 5$ je vektor regresní parametrů dán jako

$$[1.5, 1.15, 0.8, 0.45, 0.1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T,$$

- Pro $m = 15$ je vektor regresní parametrů dán jako

$$[1.5, 1.4, 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0, 0, 0, 0, 0]^T.$$


m = 5
1.5
1.0
0.5
0.0
β_j
5
10
15
20
j


m = 15
1.5
1.0
0.5
0.0
β_j
5
10
15
20
j

Shrnutí nastavení Monte Carlo simulace

- $\alpha_{remove} = 0.05$.
- $\sigma^2 = 6.25$.
- $k = 20$.
- Dvě hodnoty n ($n = 50$, $n = 300$), dvě hodnoty ρ ($\rho = 0$, $\rho = 0.8$) a dvě hodnoty m ($m = 5$, $m = 15$). Celkem $2^3 = 8$ možných nastavení.
- Počet simulací pro každé nastavení: 1000.
  - směrodatné chyby všech relativních četností jsou omezeny shora hodnotou 0.016.
  - odhady směrodatných chyb různých výběrových průměrů (viz níže) nepřesahují 0.06.

1 Metody výběru proměnných do modelu

2 Monte Carlo simulace, backward metoda

3 Pravděpodobnosti různých finálních modelů

4 Inference

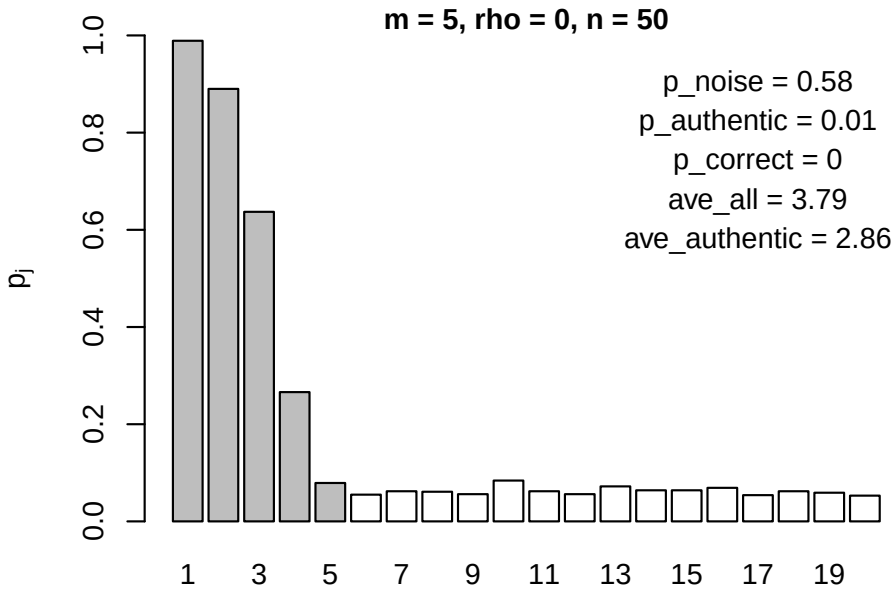
5 Přesnost odhadu regresní funkce

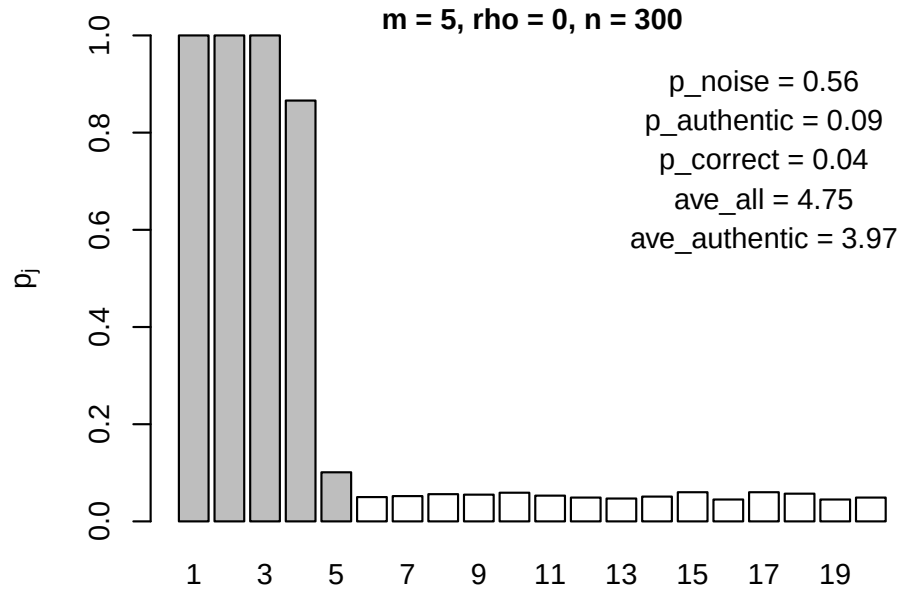
6 Rozšíření výsledků

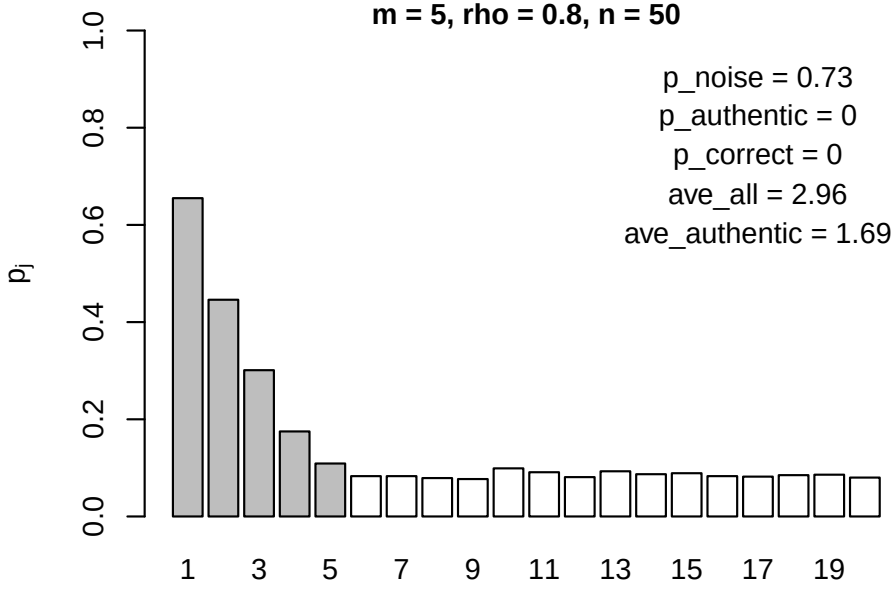
Značení

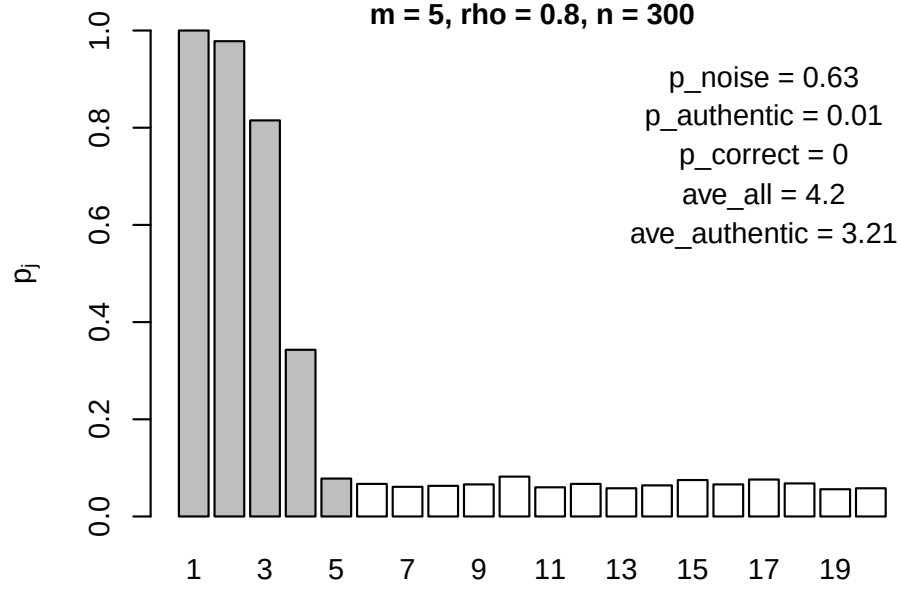
- p_{noise}: Relativní četnost toho, že ve výsledném modelu zůstane alespoň jedna vysvětlující proměnná, jejíž skutečný regresní parametr je nulový.
- $p_{authentic}$: Relativní četnost toho, že ve výsledném modelu budou všechny vysvětlující proměnné, jejichž skutečné parametry jsou nenulové.
- $p_{correct}$: Relativní četnost toho, že ve výsledném modelu budou všechny vysvětlující proměnné, jejichž skutečné parametry jsou nenulové, a nebude tam žádná vysvětlující proměnná, jejíž skutečný parametr je nulový.
- p_j (pro $j = 1, ..., k$): Relativní četnost toho, že j-tá vysvětlující proměnná bude přítomna ve finálním modelu.

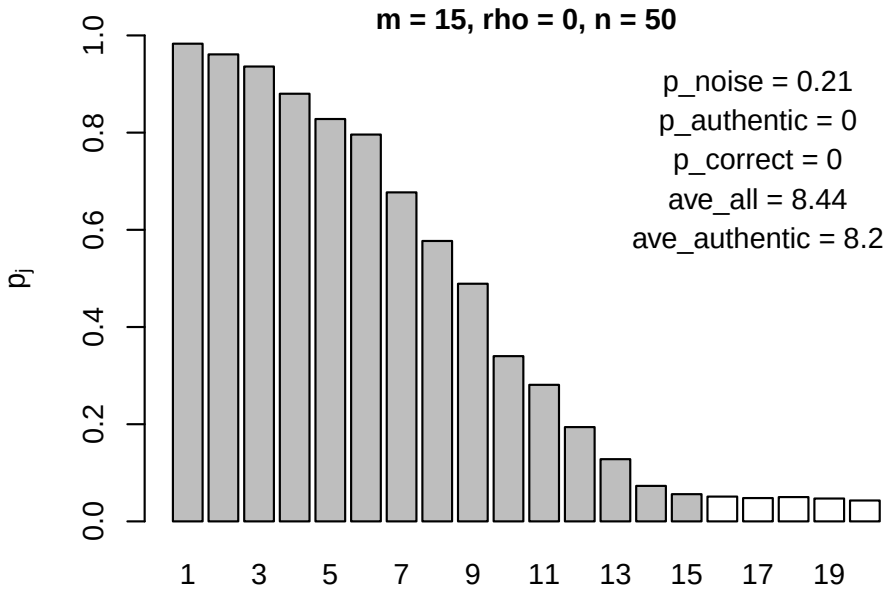
- ave_{all} : Průměrný počet všech vysvětlujících proměnných ve finálním modelu.
- $ave_{authentic}$: Průměrný počet vysvětlujících proměnných ve finálním modelu, jejichž skutečné parametry jsou nenulové.

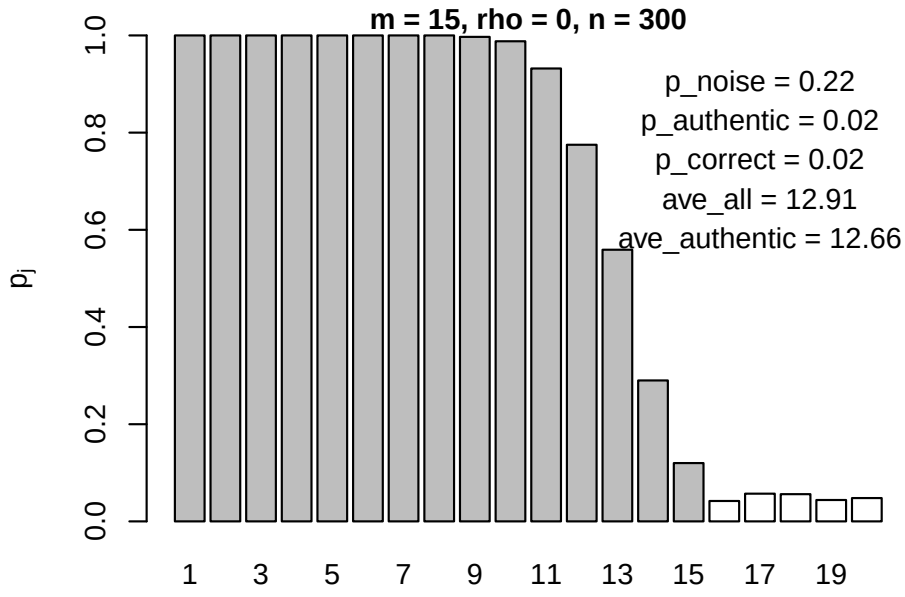

m = 5, rho = 0, n = 50
p_noise = 0.58
p_authentic = 0.01
p_correct = 0
ave_all = 3.79
ave_authentic = 2.86
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
13
15
17
19

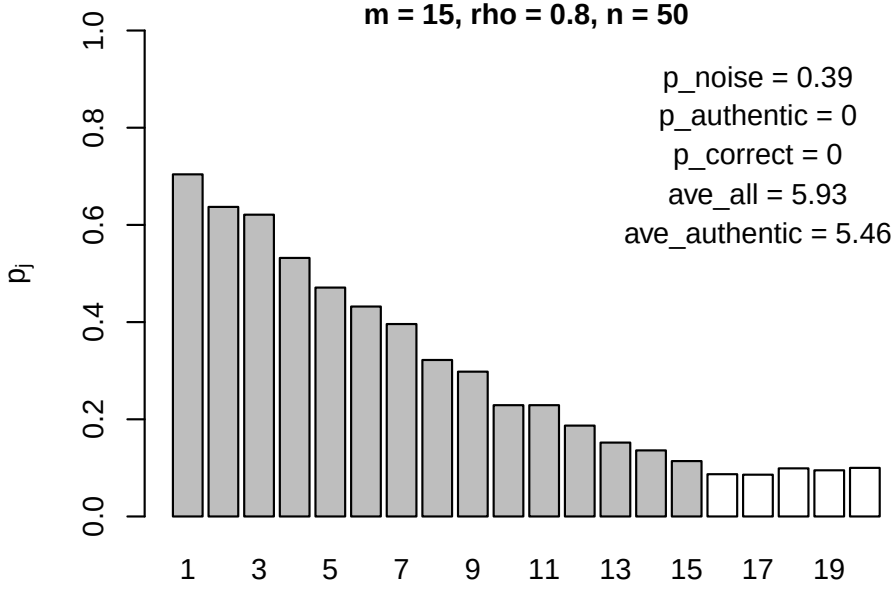

m = 5, rho = 0, n = 300
p_noise = 0.56
p_authentic = 0.09
p_correct = 0.04
ave_all = 4.75
ave_authentic = 3.97
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
13
15
17
19

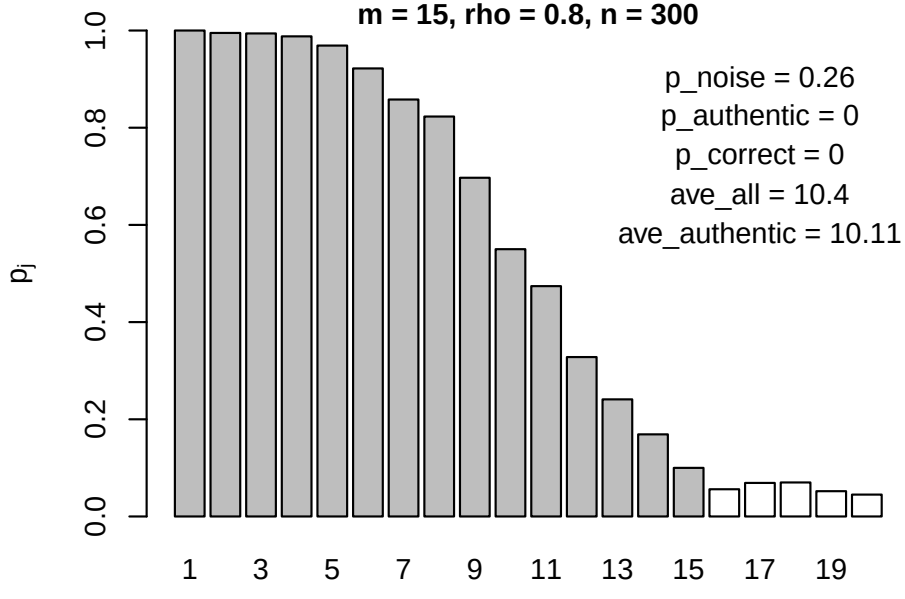

m = 5, rho = 0.8, n = 50
p_noise = 0.73
p_authentic = 0
p_correct = 0
ave_all = 2.96
ave_authentic = 1.69
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
13
15
17
19


m = 5, rho = 0.8, n = 300
p_noise = 0.63
p_authentic = 0.01
p_correct = 0
ave_all = 4.2
ave_authentic = 3.21
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
13
15
17
19


m = 15, rho = 0, n = 50
p_noise = 0.21
p_authentic = 0
p_correct = 0
ave_all = 8.44
ave_authentic = 8.2
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1 3 5 7 9 11 13 15 17 19


m = 15, rho = 0, n = 300
p_noise = 0.22
p_authentic = 0.02
p_correct = 0.02
ave_all = 12.91
ave_authentic = 12.66
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1 3 5 7 9 11 13 15 17 19


m = 15, rho = 0.8, n = 50
p_noise = 0.39
p_authentic = 0
p_correct = 0
ave_all = 5.93
ave_authentic = 5.46
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1 3 5 7 9 11 13 15 17 19


m = 15, rho = 0.8, n = 300
p_noise = 0.26
p_authentic = 0
p_correct = 0
ave_all = 10.4
ave_authentic = 10.11
p_j
0.0
0.2
0.4
0.6
0.8
1.0
1 3 5 7 9 11 13 15 17 19

Shrnutí "Pravděpodobnosti různých finálních modelů"

- $p_{correct}$ a $p_{authentic}$ jsou nízké, p_{noise} je relativně vysoká.
- Vyšší hodnota m vede k nižší p_{noise} a $p_{authentic}$.
- Vyšší hodnota ρ vede k vyšší p_{noise} a nižší $p_{authentic}$ a $p_{correct}$.
- Vyšší hodnota n vede k mírně vyšší $p_{authentic}$ a $p_{correct}$.
- Metoda *forward* vykazuje kvalitativně podobné chování.

1 Metody výběru proměnných do modelu

2 Monte Carlo simulace, backward metoda

3 Pravděpodobnosti různých finálních modelů

4 Inference

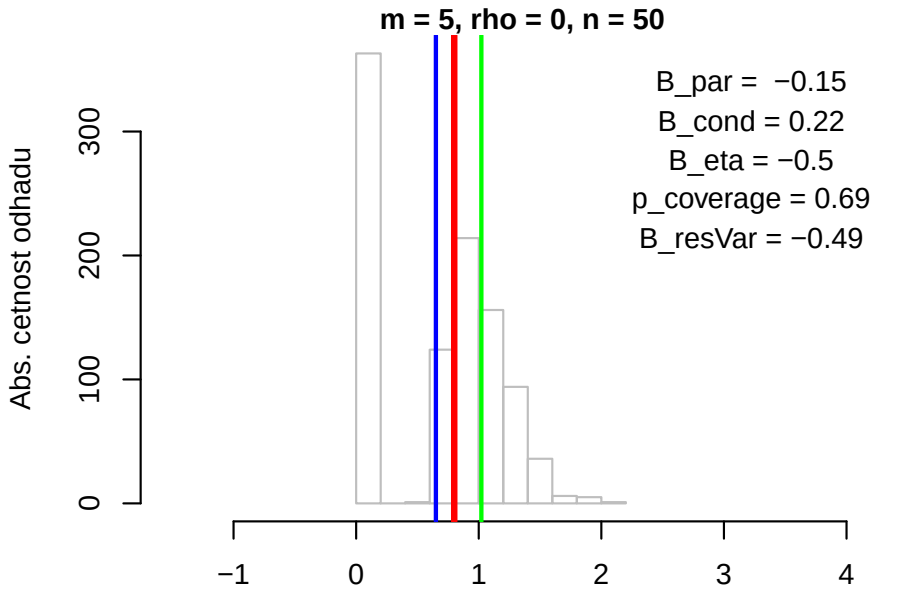
5 Přesnost odhadu regresní funkce

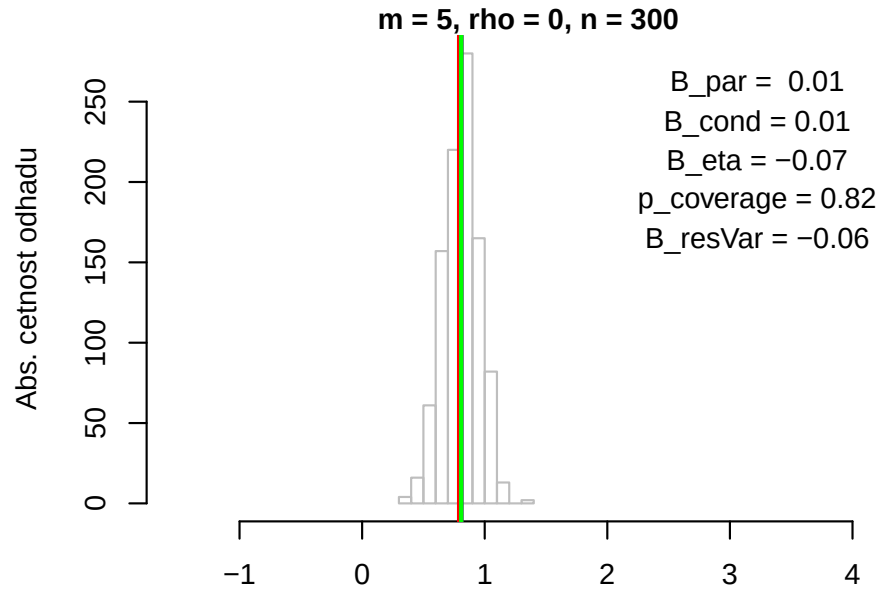
6 Rozšíření výsledků

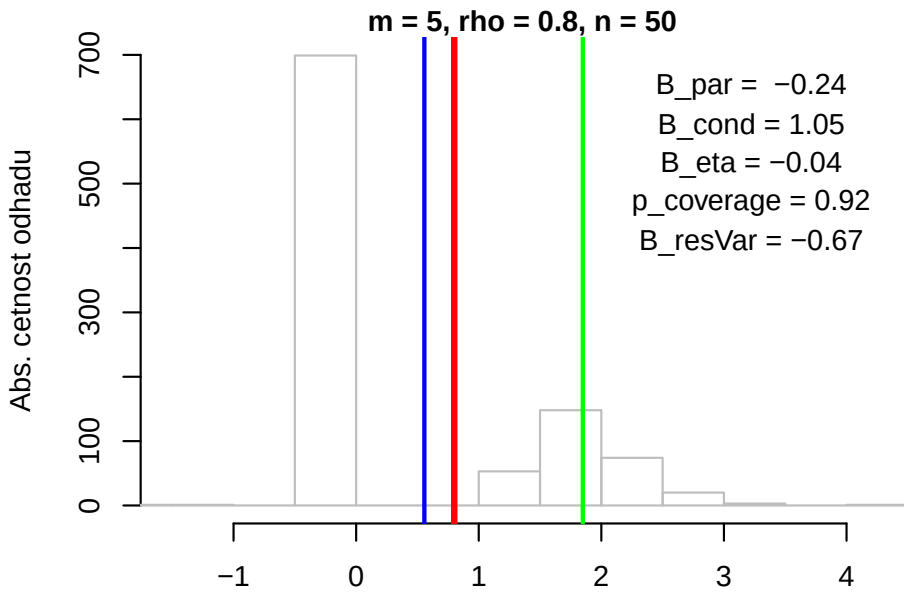
Značení

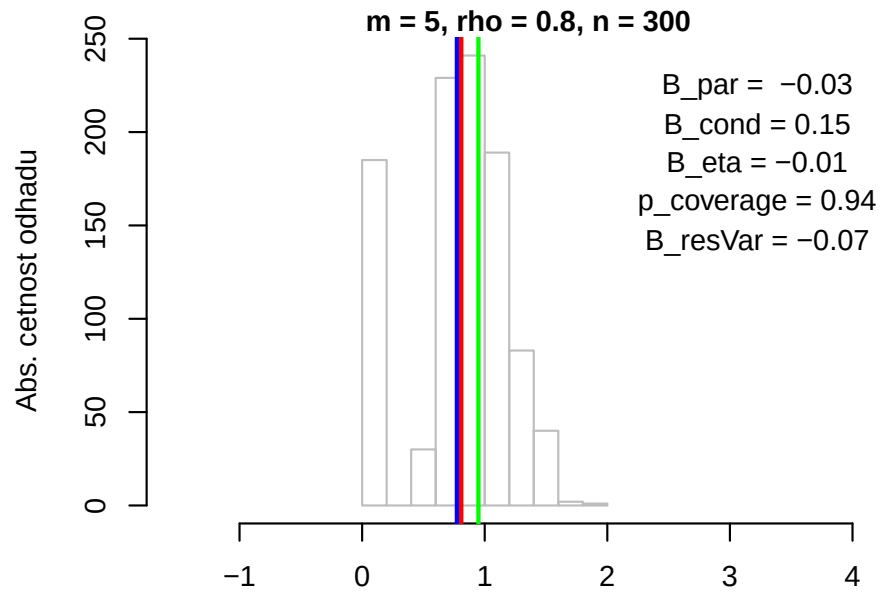
- B_{par}: Odhad vychýlení odhadu regresního parametru $\beta_3 = 0.8$ (pro $m = 5$) resp. $\beta_8 = 0.8$ (pro $m = 15$).
- B_{cond}: Odhad podmíněného vychýlení odhadu regresního parametru $\beta_3 = 0.8$ (pro $m = 5$) resp. $\beta_8 = 0.8$ (pro $m = 15$) za podmínky, že se příslušná vysvětlující proměnná vyskytuje ve finálním modelu.
- B_{eta}: Odhad vychýlení odhadu regresní funkce v bodě $[1, 1, ..., 1]$. Skutečná regresní funkce má v tomto bodě hodnotu 4 (pro $m = 5$) resp. 12 (pro $m = 15$).

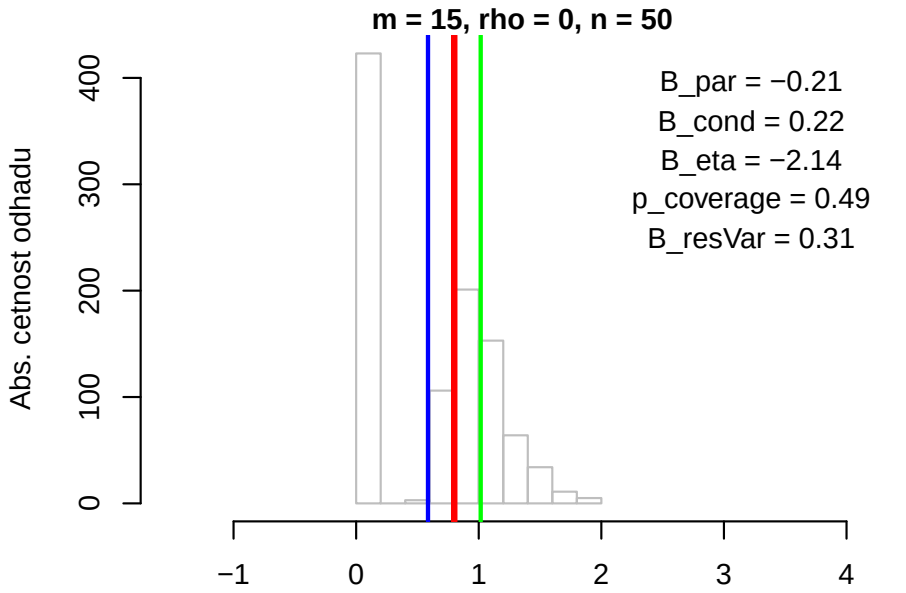
- $p_{coverage}$: Relativní četnost pokrytí regresní funkce v bodě $[1, 1, ..., 1]$ 95% intervalem spolehlivosti pro regresní funkci napočítáným dle standardní teorie na základě finálního modelu.
- B_{resVar}: Odhad vychýlení odhadu rozptylu chybové složky prostřednictvím reziduálního rozptylu napočteného na základě finálního modelu.

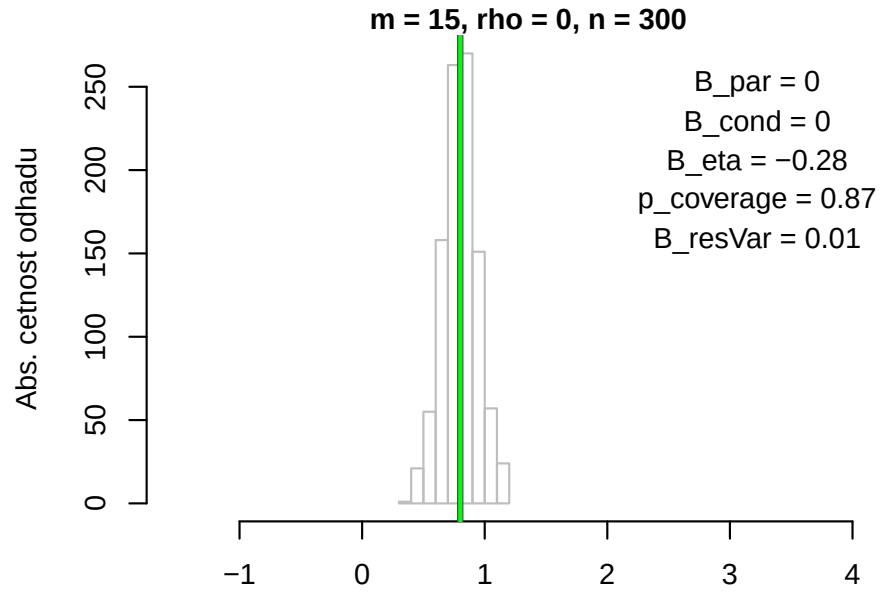

m = 5, rho = 0, n = 50
B_par = −0.15
B_cond = 0.22
B_eta = −0.5
p_coverage = 0.69
B_resVar = −0.49
Abs. cetnost odhadu
0
100
200
300
−1
0
1
2
3
4

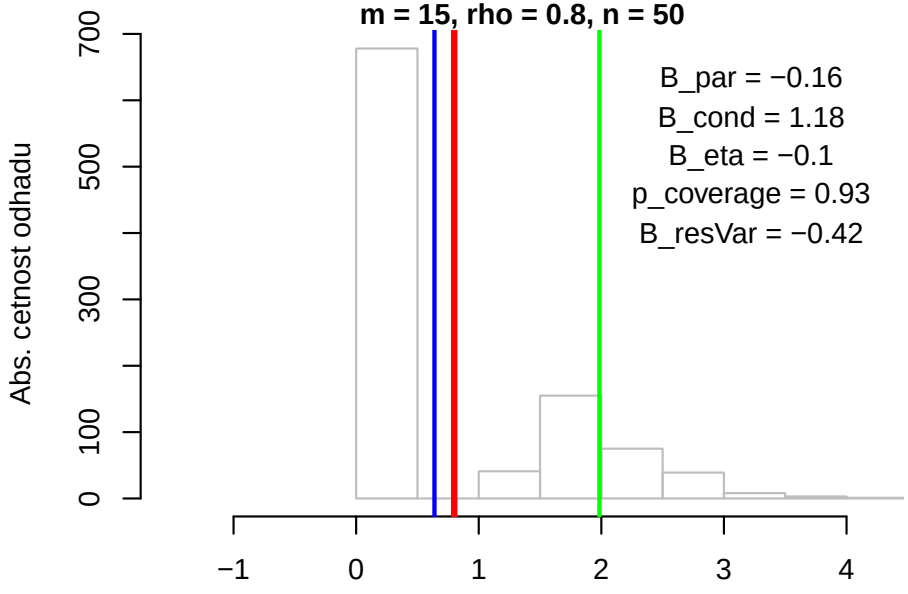

m = 5, rho = 0, n = 300
B_par = 0.01
B_cond = 0.01
B_eta = −0.07
p_coverage = 0.82
B_resVar = −0.06
Abs. cetnost odhadu
0
50
100
150
200
250
−1
0
1
2
3
4

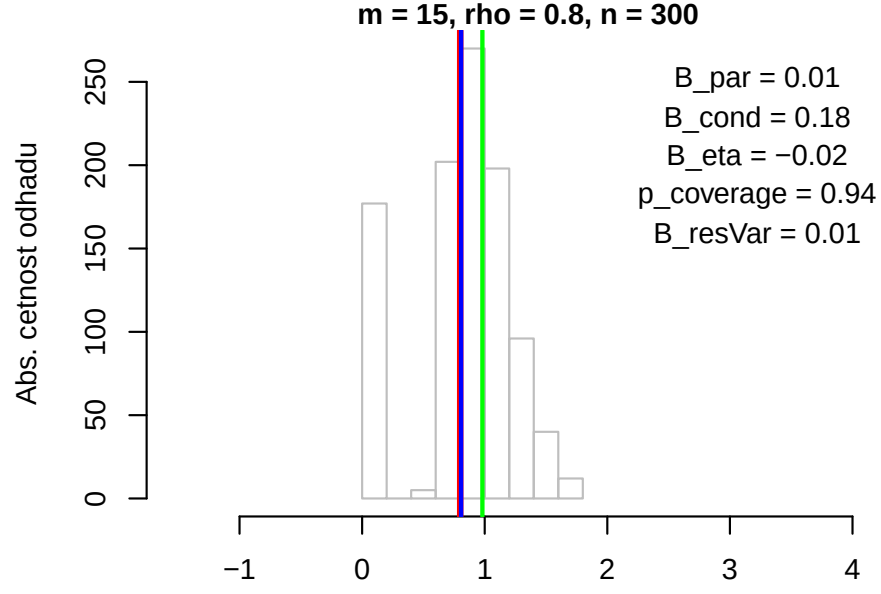

m = 5, rho = 0.8, n = 50
B_par = −0.24
B_cond = 1.05
B_eta = −0.04
p_coverage = 0.92
B_resVar = −0.67
Abs. cetnost odhadu
0
100
300
500
700
−1
0
1
2
3
4


m = 5, rho = 0.8, n = 300
B_par = −0.03
B_cond = 0.15
B_eta = −0.01
p_coverage = 0.94
B_resVar = −0.07
Abs. cetnost odhadu
0
50
100
150
200
250
−1
0
1
2
3
4


m = 15, rho = 0, n = 50
B_par = −0.21
B_cond = 0.22
B_eta = −2.14
p_coverage = 0.49
B_resVar = 0.31
Abs. cetnost odhadu
0
100
200
300
400
−1
0
1
2
3
4


m = 15, rho = 0, n = 300
B_par = 0
B_cond = 0
B_eta = −0.28
p_coverage = 0.87
B_resVar = 0.01
Abs. cetnost odhadu
0
50
100
150
200
250
−1
0
1
2
3
4


m = 15, rho = 0.8, n = 50
B_par = −0.16
B_cond = 1.18
B_eta = −0.1
p_coverage = 0.93
B_resVar = −0.42
Abs. cetnost odhadu
0
100
300
500
700
−1
0
1
2
3
4


m = 15, rho = 0.8, n = 300
B_par = 0.01
B_cond = 0.18
B_eta = −0.02
p_coverage = 0.94
B_resVar = 0.01
Abs. cetnost odhadu
0
50
100
150
200
250
−1
0
1
2
3
4

Shrnutí "Inference"

- Odhady regresních parametrů jsou vychýlené.
- Pokud je vysvětlující proměnná zahrnuta ve finálním modelu, tak je příslusný odhadnutý efekt v průměru nadhodnocený.
- Odhady regresní funkce jsou v obecnosti vychýlené a intervaly spolehlivosti pro regresní funkci nemají požadovaná pokrytí.
- Reziduální rozptyl napočtený z finálního modelu dle standardní teorie je vychýleným odhadem rozptylu chybové složky.
- Metoda *forward* vykazuje kvalitativně podobné chování.

1 Metody výběru proměnných do modelu

2 Monte Carlo simulace, backward metoda

3 Pravděpodobnosti různých finálních modelů

4 Inference

5 Přesnost odhadu regresní funkce

6 Rozšíření výsledků

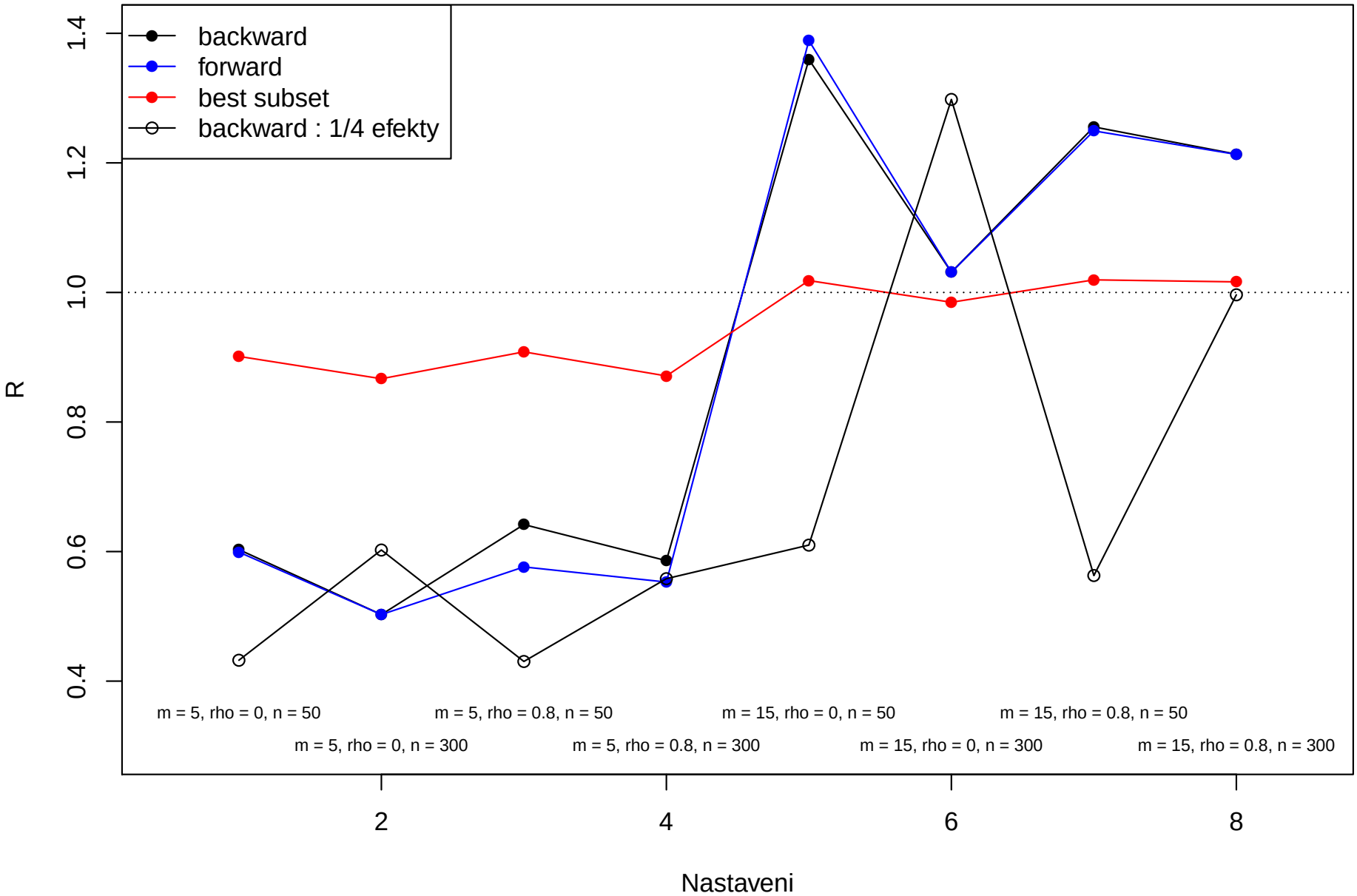
Přesnost odhadu regresní funkce

- S využitím Monte Carlo simulace odhadujeme

$$R_{backward} = \frac{\overline{MSE}_{backward}}{\overline{MSE}_{full}},$$

kde $\overline{MSE}_{backward}$ resp. $\overline{MSE}_{full} = \sigma^2 \frac{k}{n}$ jsou průměrné (napříč n původními body) střední čtvercové chyby odhadu regresní funkce na základě finálního modelu z krokové metody backward a na základě plného modelu s k vysvětlujícími proměnnými.

- Výsledky zakreslíme pro 8 nastavení a porovnáme je s $R_{forward}$ (pro $\alpha_{add} = 0.05$), s $R_{bestSubset}$ založené na upraveném indexu determinace a s metodou *backward* v situaci, kdy regresní parametry mají jen *čtvrtinovou velikost* oproti původnímu nastavení.


backward
forward
best subset
backward : 1/4 efekty
R
0.4
0.6
0.8
1.0
1.2
1.4
m = 5, rho = 0, n = 50
m = 5, rho = 0.8, n = 50
m = 15, rho = 0, n = 50
m = 15, rho = 0.8, n = 50
m = 5, rho = 0, n = 300
m = 5, rho = 0.8, n = 300
m = 15, rho = 0, n = 300
m = 15, rho = 0.8, n = 300
2
4
6
8
Nastaveni

Shrnutí “Přesnosti odhadu regresní funkce”

- Pro m=5 poskytla kroková metoda přesnější odhad regresní funkce (měřeno MSE) a potažmo přesnější předpovědi než plný model.
- Metoda best subset není záchranou.

1 Metody výběru proměnných do modelu

2 Monte Carlo simulace, backward metoda

3 Pravděpodobnosti různých finálních modelů

4 Inference

5 Přesnost odhadu regresní funkce

6 Rozšíření výsledků

- Domníváme se, že analogické výsledky lze očekávat i v případě zobecněných lineárních modelů.