


## Semestrální projekt

Tato domácí úloha byla zpracována pomocí softwaru  a zabývá se datovým souborem `SaratogaHouses` z balíku `mosaicData`. Cílem analýzy bylo vytvořit model, který by popisoval závislost vysvětlované proměnné `price`, jež obsahuje informaci o ceně nemovitostí za rok 2006 v Saratoga County (New York, USA) v dolarech, na několika vybraných vysvětlujících proměnných. Datový soubor musel být nejprve náhodně rozdělen na data trénovací a testovací. Na trénovacích datech byly provedeny všechny zadané podúkoly včetně výběru finálního modelu, jehož kvalita byla následně prověřena právě na datech testovacích. Rozdělení dat na trénovací a testovací proběhlo na základě sekvence příkazů přiložených v zadání tohoto úkolu. Po zběžném náhledu dat, který neprokázal žádné zásadní problémy v proměnných, již bylo přikročeno k jednotlivým podúkolům.

### Fáze 1

Prvním úkolem bylo najít kvantitativní vysvětlující proměnnou, která vykazuje nejvyšší hodnotu indexu determinace v příslušném modelu s proměnnou `price`. Poté byl stejný úkol proveden ještě při posuzování na základě kvadrátu výběrového korelačního koeficientu. Hodnoty indexů determinace a výběrových korelačních koeficientů byly pro každý model shodné, což není náhoda. V jednoduchém regresním modelu jsou totiž tyto dvě hodnoty vždy totožné.

Kvantitativní vysvětlující proměnnou vykazující nejvyšší hodnotu indexu determinace a zároveň nejvyšší hodnotu kvadrátu výběrového korelačního koeficientu je proměnná `livingArea`, jež obsahuje rozměry obytné části domu ve stopách čtverečních. Výsledek této fáze úlohy je logický – bylo možné předpokládat, že cena domu bude velmi záviset na rozloze jeho obytné části. Je tudíž zřejmé, že právě tato proměnná bude nepochybně hrát ve finálním modelu významnou roli.

```
determinace = data.frame(r11,r12,r13,r14,r15,r16,r17,r18,r19); determinace

##           r11           r12           r13           r14           r15           r16           r17
## 1 0.01766417 0.03039657 0.3218979 0.5132662 0.04074822 0.1476238 0.150209
##           r18           r19
## 1 0.3532691 0.2897921

which.max(determinace)

## r14
## 4

max(determinace)

## [1] 0.5132662

#nejvyssi ma tedy promenna v 5. sloupci, tj. livingArea
```

```
(korelace = data.frame(c1,c2,c3,c4,c5,c6,c7,c8,c9))

##           c1           c2           c3           c4           c5           c6           c7
## 1 0.01766417 0.03039657 0.3218979 0.5132662 0.04074822 0.1476238 0.150209
##           c8           c9
## 1 0.3532691 0.2897921

which.max(korelace)

## c4
## 4

max(korelace)

## [1] 0.5132662

#stejný výsledek :) nejuv kvadrat vyber korel koeficientu je opet s promennou livingArea
```

## Fáze 2

Tato fáze úkolu se podrobněji zabývá závislostí ceny nemovitosti na vysvětlující proměnné bedrooms, která informuje o počtu ložnic v domě. Jak výstup ze softwaru ukazuje, odhad regresního koeficientu u této proměnné vyšel v jednoduchém regresním modelu kladně, což je logické vzhledem k poměrně silné kladné lineární závislosti mezi cenou a počtem ložnic. Zásadní změna u odhadu tohoto regresního koeficientu ovšem nastává v situaci, kdy byla do tohoto modelu přidána jako další vysvětlující proměnná již zmiňovaná livingArea. Výstup z druhého modelu ukázal změnu znaménka u odhadu regresního koeficientu pro proměnnou bedrooms.

```
#faze 2
summary(m16) #price~berooms - cim vic loznic, tim vyssi cena za byt


##
## Call:
## lm(formula = price ~ train[, 7])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -264787  -57542  -19518   33633   567638
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    66043      9943    6.642 4.43e-11 ***
## train[, 7]     47106      3047   15.460 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93470 on 1380 degrees of freedom
## Multiple R-squared:  0.1476, Adjusted R-squared:  0.147
```

```

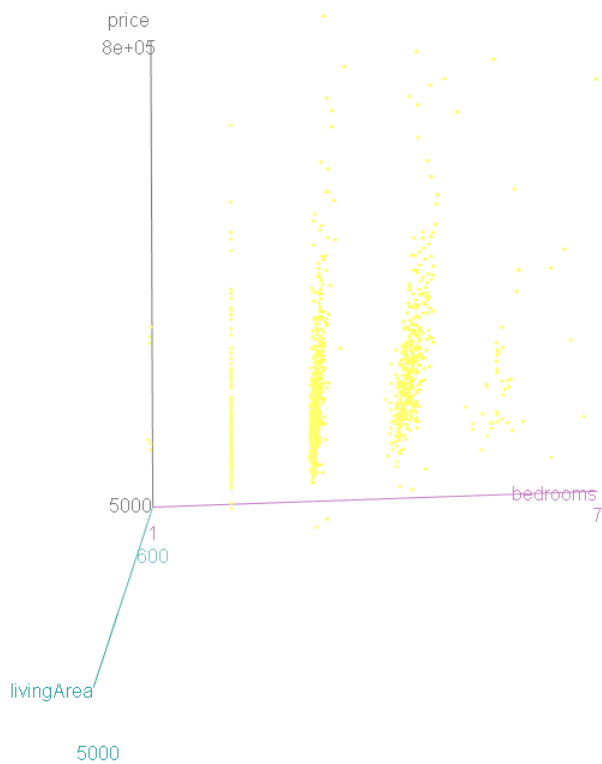
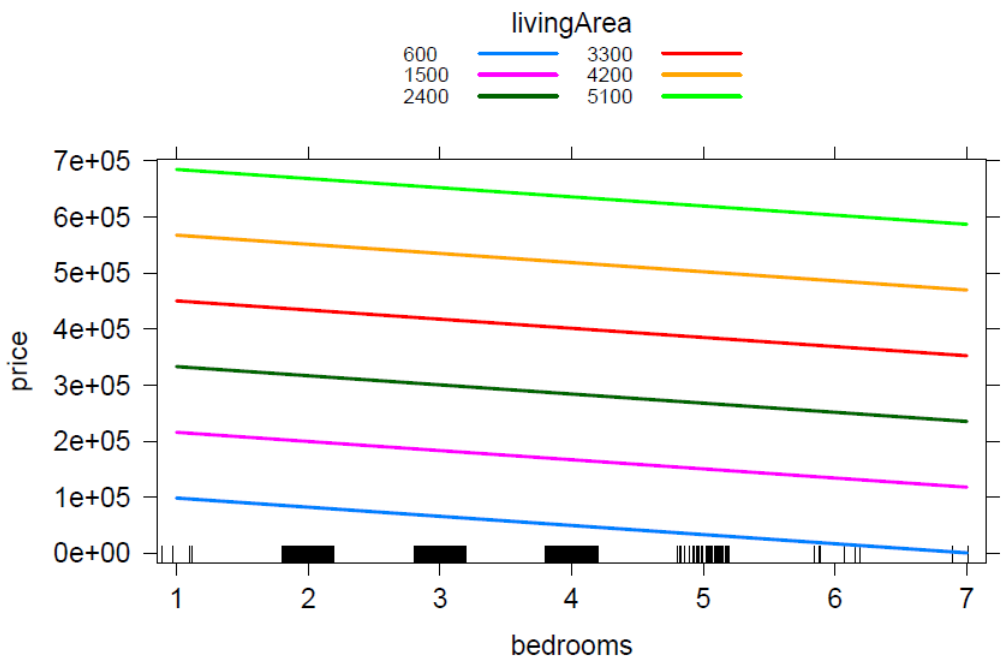
m21 = lm(price~ bedrooms + livingArea); summary(m21) #nyni je u promenne bedrooms

##
## Call:
## lm(formula = price ~ bedrooms + livingArea)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -290982  -41498   -9089   27329  549501
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36477.35     7489.57   4.870 1.24e-06 ***
## bedrooms    -16315.68     2981.36  -5.473 5.26e-08 ***
## livingArea    130.32         3.95  32.991 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69900 on 1379 degrees of freedom
## Multiple R-squared:  0.5236, Adjusted R-squared:  0.5229
## F-statistic: 757.9 on 2 and 1379 DF,  p-value: < 2.2e-16

```

Nejpravděpodobnějším vysvětlením tohoto jevu bude nejspíš fakt, že při stejné rozloze obytných prostor je větší poptávka po domech s menším množstvím pokojů (které jsou tudíž větší a prostornější) oproti stejně velkým domům s velkým množstvím malých pokojů. Samozřejmě obecně platí, že čím víc pokojů, tím lépe, ale to především proto, že předpokládáme, že čím více pokojů, tím větší dům. Proto je odhad koeficientu v prvním z modelů kladný a v situaci, kdy je brána v potaz také rozloha, již nikoli. Tuto myšlenku, která vysvětluje rozpor modelu jednoduché a vícenásobné lineární regrese vzhledem k proměnné bedrooms, dokládají také některé následující výstupy z .

bedrooms\*livingArea effect plot



### Fáze 3

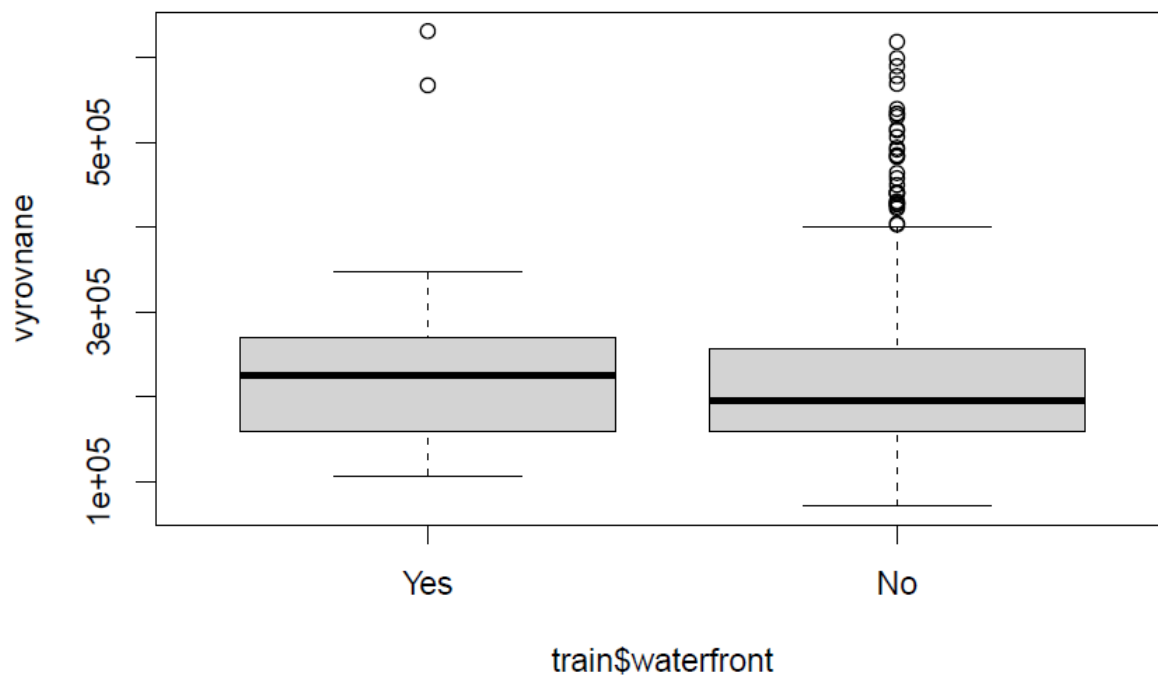
Ve třetí fázi úkolu byl vytvořen model vícenásobné lineární regrese popisující závislost ceny nemovitosti na všech ostatních kvantitativních proměnných. Odhad regresních koeficientů tohoto modelu lze vidět níže.

```
#faze 3 - model kvantitativnich promennych
model_kvanti = lm(price~ lotSize + age + landValue + livingArea
                  + pctCollege + bedrooms + fireplaces + bathrooms + rooms)
summary(model_kvanti)

##
## Call:
## lm(formula = price ~ lotSize + age + landValue + livingArea +
##     pctCollege + bedrooms + fireplaces + bathrooms + rooms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -225785  -37125   -6373    27588   456161
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.186e+04  1.132e+04   1.047  0.29524
## lotSize       6.593e+03  2.597e+03   2.539  0.01123 *
## age          -9.105e+01  6.225e+01  -1.463  0.14376
## landValue     9.024e-01  5.192e-02  17.379 < 2e-16 ***
## livingArea    7.343e+01  5.507e+00  13.334 < 2e-16 ***
## pctCollege    1.474e+02  1.716e+02   0.859  0.39038
## bedrooms     -1.255e+04  2.946e+03  -4.261  2.17e-05 ***
## fireplaces    4.391e+03  3.522e+03   1.247  0.21269
## bathrooms    2.347e+04  3.943e+03   5.952  3.35e-09 ***
## rooms        3.547e+03  1.152e+03   3.080  0.00211 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62130 on 1372 degrees of freedom
## Multiple R-squared:  0.6256, Adjusted R-squared:  0.6231
## F-statistic: 254.7 on 9 and 1372 DF, p-value: < 2.2e-16

# Vyrovnane hodnoty
vyrovnane=fitted(model_kvanti)
```

Vyrovnané hodnoty tohoto modelu byly následně graficky zobrazeny pomocí boxplotů vytvořených odděleně pro kategorie (Yes a No) kategoriální proměnné waterfront, která uvádí, zda se nemovitost nalézá u vodní plochy či nikoliv.

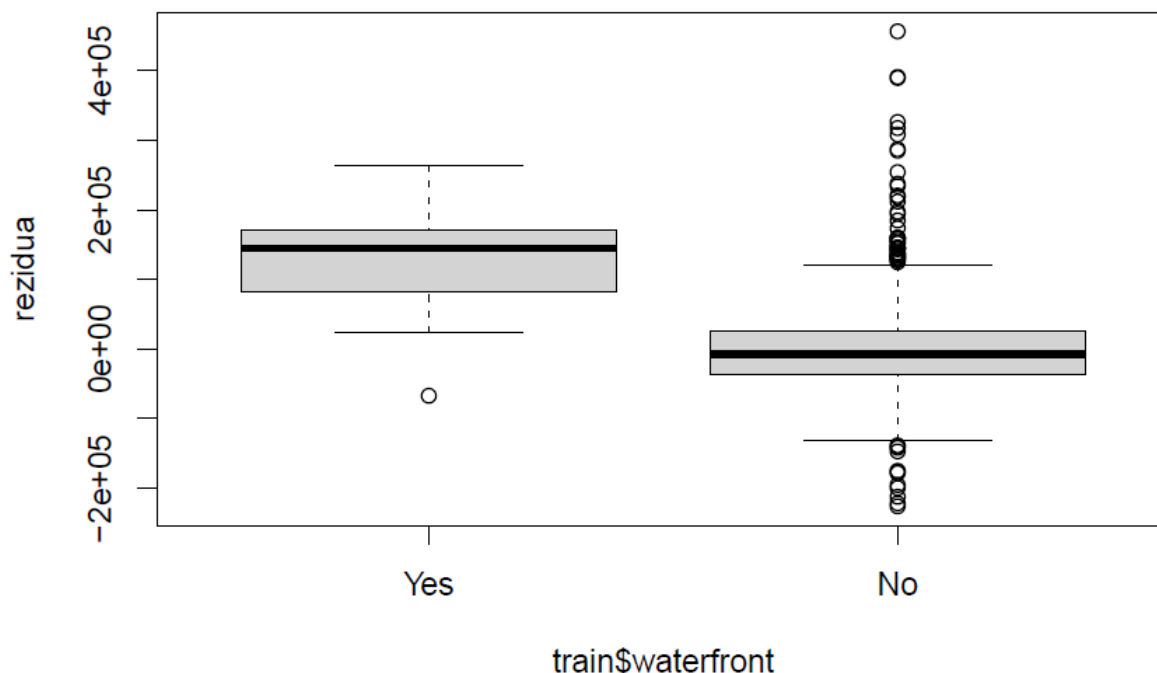


Pro tento model byla poté vypočítána průměrná hodnota reziduí obou kategorií (Yes a No) zvlášť a byly vytvořeny boxploty reziduí (opět pro každou kategorii zvlášť).

```
#rezidua:
rezidua = resid(model_kvanti)
aggregate(rezidua, list(train$waterfront), FUN=mean)
```

```
##   Group.1      x
## 1    Yes 124620.823
## 2    No -1183.397
```

```
boxplot(rezidua~train$waterfront)
```



Dle posledních dvou boxplotů je možné říci, že rezidua jsou v modelu kvantitativních proměnných nadhodnocena pro kategorii Yes a naopak lehce podhodnocena u kategorie No. To znamená, že tento model podhodnocuje cenu nemovitostí nacházejících se u vodních ploch a zároveň nadhodnocuje ceny domů daleko od vodních ploch. Z těchto úsudků vyplývá, že proměnná `waterfront` bude hrát nejspíš v odhadu ceny nemovitosti významnou roli a bude vhodné ji ve finálním modelu uvažovat.

#### Fáze 4

Úkolem ve čtvrté fázi bylo sestavit model vícenásobné lineární regrese zachycující závislost proměnné `price` na všech kvantitativních a také kategoriálních vysvětlujících proměnných, prozatím bez interakcí a nelinearity. Kategoriální proměnné vystupují v modelu v podobě dummy proměnných. Vzhledem k tomu, že všechny kategoriální proměnné byly již předem nastaveny na typ `factor`, nebylo nutné je k tomuto účelu nijak dále modifikovat.

Dalším úkolem bylo interpretovat odhad regresního parametru u dummy proměnné reprezentující kategoriální proměnnou `centralAir`, tj. přítomnost klimatizace.

```
# Model se zařazením kategoriální proměnné
model_all = lm(price ~ lotSize + age + landValue + livingArea + pctCollege + bed
coef(summary(model_all))[19,]

##      Estimate      Std. Error      t value      Pr(>|t|)
## -1.043690e+04  4.020213e+03 -2.596107e+00  9.530044e-03

#zjisteni referencni hodnoty promenne centralAir:
contrasts(train[,16]) #referencni hodnota je Yes

##      No
## Yes   0
## No    1
```

Na základě výstupu zobrazeného výše lze konstatovat, že při jinak stejných podmínkách je odhad střední hodnoty ceny nemovitosti přibližně o 10 437 dolarů nižší pro nemovitost bez klimatizace oproti nemovitosti s klimatizací. Tento závěr se zhruba shoduje s mou původní představou.

Dále byla uvažována nejprve ta pozorování, pro něž přítomnost klimatizace nabývá hodnoty Yes. S využitím těchto pozorování byl opět sestaven model (kromě rozdílného využití údajů o (ne)přítomnosti klimatizace se model neliší od předchozího). Stejný model byl sestaven rovněž pro nemovitosti bez klimatizace, a následně byly odhady regresních koeficientů u těchto dvou modelů vzájemně porovnány, aby byly vidět změny vlivů ostatních proměnných na cenu nemovitostí s a bez klimatizace.

```
klima_yes_data = subset(train, train$centralAir=="Yes")
model_all_klima = lm(price ~ lotSize + age + landValue + livingArea + pctCollege + bedrooms
+ fireplaces + bathrooms + rooms + heating + fuel + sewer + waterfront
+ newConstruction, data=klima_yes_data)
#summary(model_all_klima)

klima_no_data = subset(train, train$centralAir=="No")
model_all_No_klima = lm(price ~ lotSize + age + landValue + livingArea + pctCollege + bedrooms
+ fireplaces + bathrooms + rooms + heating + fuel + sewer + waterfront
+ newConstruction, data=klima_no_data)
#summary(model_all_No_klima)

compareCoefs(model_all_klima,model_all_No_klima)

## Calls:
## 1: lm(formula = price ~ lotSize + age + landValue + livingArea + pctCollege
## + bedrooms + fireplaces + bathrooms + rooms + heating + fuel + sewer +
## waterfront + newConstruction, data = klima_yes_data)
## 2: lm(formula = price ~ lotSize + age + landValue + livingArea + pctCollege
## + bedrooms + fireplaces + bathrooms + rooms + heating + fuel + sewer +
## waterfront + newConstruction, data = klima_no_data)
```



##	Model 1	Model 2
## (Intercept)	75792	143678
## SE	45559	23841
##		
## lotSize	2019	9493
## SE	5661	2982
##		
## age	-547.8	-84.7
## SE	270.1	62.7
##		
## landValue	0.7135	0.9375
## SE	0.0838	0.0712
##		
## livingArea	87.35	61.10
## SE	10.57	5.99
##		
## pctCollege	-331	156
## SE	434	177
##		
## bedrooms	-16246	-3474
## SE	5574	3235
##		
## fireplaces	4706	1801
## SE	6532	3934
##		
## bathrooms	28738	16218
## SE	7243	4407
##		
## rooms	5452	2417
## SE	2079	1254
##		
## heatinghot water/steam	-16874	-7526
## SE	15080	4629
##		
## heatingelectric	5203	4639
## SE	22240	21835
##		
## fuelelectric	-27653	-14184
## SE	20565	21738
##		
## fueloil	13142	-9268
## SE	16486	5854
##		
## sewerpublic/commercial	-4551	3181
## SE	8886	4557
##		
## sewernone	28525	-19605
## SE	32224	20374
##		
## waterfrontNo	-106622	-151407
## SE	36267	18228
##		
## newConstructionNo	63896	28737
## SE	12939	10928

Vidíme, že v případě modelu s klimatizovanými nemovitostmi oproti tomu s nemovitostmi bez klimatizací roste (resp. klesá) o něco více cena nemovitosti se zvyšujícím se počtem pokojů, koupelen, krbů nebo například s rostoucí obytnou plochou. Naopak s rostoucí hodnotou proměnných *lotSize*, *age*, *landValue*, *pctCollege* nebo například *bedrooms* se více mění cena nemovitostí bez klimatizace oproti těm klimatizovaným. U některých vysvětlujících proměnných dokonce dochází ke změně znaménka u regresních parametrů oproti původnímu kompletnímu modelu.

## Fáze 5

Pokud bychom se měli pouze na základě intuice a tedy nezávisle na předchozích analýzách rozhodnout, jaké vysvětlující proměnné do modelu zahrnout, nepochybně by to byla proměnná *livingArea*. Další vhodnou vysvětlující proměnnou by mohla být například hodnota půdy v dolarech (*landValue*), jelikož by mělo platit, že čím hodnotnější pozemek, tím vyšší cena nemovitosti. Tuto proměnnou do zkušebního modelu zkusíme zapojit ve formě polynomu druhého stupně, jelikož jsem toho názoru, že bude cena stoupat s hodnotou půdy „rychleji“ než lineárně. Jako kategoriální proměnná vstoupí do modelu *newConstruction* s informací o tom, zda se jedná o novostavbu či nikoli, protože si myslím, že je možné, že i tento údaj by mohl mít na finální cenu vliv. Poslední proměnnou ve zkušebním modelu bude počet ložnic – ten vložíme do modelu zároveň v interakci s obytnou plochou, jelikož jsme již v předchozích fázích uvedli úvahu o vzájemného působení obytné plochy a počtu ložnic (tato úvaha proběhla již před samotnou analýzou, a je proto možné ji v tomto intuitivním modelu zohlednit).

Vzniklý zkušební model lze tedy zapsat jako

$$price_i = \beta_0 + \beta_1 \cdot landValue_i + \beta_2 \cdot landValue_i^2 + \beta_3 \cdot livingArea_i + \beta_4 \cdot newConstruction_i + \beta_5 \cdot bedrooms_i + \beta_6 \cdot livingArea_i \cdot bedrooms_i + \varepsilon_i,$$

přičemž koeficient  $\beta_0$  lze interpretovat jako střední hodnotu ceny nemovitosti při nulové hodnotě všech vysvětlujících proměnných,  $\beta_3 + \beta_6 \cdot bedrooms_i$  jako změnu střední hodnoty ceny domu při vzrůstu obytné plochy o 1 *ceteris paribus*,  $\beta_4$  jako změnu střední hodnoty ceny staršího domu oproti novostavbě *ceteris paribus* (ref. hodnota je Yes),  $\beta_5 + \beta_6 \cdot livingArea_i$  je změna střední hodnoty ceny domu při zvýšení počtu ložnic o 1 *ceteris paribus*. Koeficienty  $\beta_1$  a  $\beta_2$  jsou vzhledem k polynomu obtížně interpretovatelné.

Shrnutí odhadu takového modelu je k nahlédnutí níže.

```
summary(model_vlastni)

##
## Call:
## lm(formula = price ~ poly(landValue, degree = 2, raw = TRUE) +
##     livingArea + newConstruction + bedrooms + livingArea:bedrooms,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -262582  -37802   -5102    28181   468740
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -4.206e+04  1.962e+04  -2.144   0.0322
## poly(landValue, degree = 2, raw = TRUE)1  1.093e+00  1.010e-01  10.820 < 2e-16
## poly(landValue, degree = 2, raw = TRUE)2 -8.840e-07  4.475e-07  -1.976   0.0484
## livingArea                        1.245e+02  1.086e+01  11.457 < 2e-16
## newConstructionNo                 4.171e+04  8.195e+03   5.089  4.1e-07
## bedrooms                       -2.954e+03  5.575e+03  -0.530   0.5962
## livingArea:bedrooms               -4.940e+00  2.855e+00  -1.730   0.0838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62840 on 1375 degrees of freedom
## Multiple R-squared:  0.6161, Adjusted R-squared:  0.6145
## F-statistic: 367.8 on 6 and 1375 DF,  p-value: < 2.2e-16
```

Předpokládáme-li nyní pro jednoduchost splnění všech potřebných předpokladů lineárního regresního modelu, je možné nyní otestovat a vyhodnotit model. Adjustovaný koeficient determinace říká, že se zkušebním modelem podařilo zachytit 61,45 % celkové variability, což rozhodně není špatný výsledek. Dílčí t-testy u několika odhadů regresních koeficientů ale prokázaly možnou nesignifikanci vysvětlující proměnné bedrooms, model proto zřejmě nebude možné použít. Nevhodnost modelu ještě prověříme pomocí ANOVA testů. Nejprve vyzkoušíme význam polynomu v modelu. Již dílčí t-test naznačil možný problém s tímto polynomem (p-hodnota blízká 0,05), tuto nejistotu potvrzuje také příslušný ANOVA test, který jen velmi těsně zamítá na hladině významnosti 5 % nulovou hypotézu, že je odhad regresního koeficientu u druhého stupně proměnné landValue roven nule, jejíž platnost by znamenala, že je polynom této proměnné v modelu statisticky nevýznamný, a bylo by vhodné ho z modelu vyřadit.

```
model_lin = lm(price~ landValue + livingArea + newConstruction + bedrooms + livingArea:bedrooms, dat
anova(model_vlastni, model_lin) #zde je velmi tesne zamitame H0
```

```
## Analysis of Variance Table
##
## Model 1: price ~ poly(landValue, degree = 2, raw = TRUE) + livingArea +
##     newConstruction + bedrooms + livingArea:bedrooms
## Model 2: price ~ landValue + livingArea + newConstruction + bedrooms +
##     livingArea:bedrooms
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    1375 5.4291e+12
## 2    1376 5.4445e+12 -1 -1.5411e+10 3.903 0.0484 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Další ANOVA test zkoumající potřebu interakce v modelu ale již nulovou hypotézu o nevýznamnosti interakce nezamítá.

```
model_bezInt = lm(price~ poly(landValue, degree = 2, raw = TRUE) + livingArea + newConstruction +
anova(model_vlastni, model_bezInt)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ poly(landValue, degree = 2, raw = TRUE) + livingArea +
##     newConstruction + bedrooms + livingArea:bedrooms
## Model 2: price ~ poly(landValue, degree = 2, raw = TRUE) + livingArea +
##     newConstruction + bedrooms
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    1375 5.4291e+12
## 2    1376 5.4409e+12 -1 -1.182e+10 2.9936 0.08382 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#nezamitame H0: odhad bety u interakce=0 -> nezam, ze je tato interakce v modelu stat. nevyznamna!
```

Závěr z této fáze úkolu tedy je, že předpoklad o přítomnosti interakce obytné plochy s počtem ložnic v modelu byl nesprávný, dodatečný vliv vzájemného působení těchto dvou proměnných nebyl prokázán. Model bez této interakce již funguje mnohem lépe při minimální ztrátě vysvětlené variability. Nutnost polynomu druhého stupně proměnné landValue v modelu zůstává diskutabilní. Všimněme si ještě shodu p-hodnot u dílčích t-testů s odpovídajícím ANOVA F-testem, která samozřejmě není náhodou.

```

call:
lm(formula = price ~ poly(landvalue, degree = 2, raw = TRUE) +
    livingArea + newConstruction + bedrooms, data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-258056  -37674   -5701    27808   462826

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.423e+04  1.123e+04  -1.266   0.2057
poly(landvalue, degree = 2, raw = TRUE)1  1.100e+00  1.010e-01  10.890 < 2e-16
poly(landvalue, degree = 2, raw = TRUE)2 -9.015e-07  4.477e-07  -2.014   0.0442
livingArea        1.069e+02  3.930e+00  27.211 < 2e-16
newConstructionNo  4.209e+04  8.198e+03   5.134 3.25e-07
bedrooms         -1.139e+04  2.707e+03  -4.207 2.75e-05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62880 on 1376 degrees of freedom
Multiple R-squared:  0.6153,    Adjusted R-squared:  0.6139
F-statistic: 440.2 on 5 and 1376 DF,  p-value: < 2.2e-16

```

## Fáze 6

Tato fáze úkolu nadále pracuje s původním zkušebním modelem z předchozí fáze a úkolem bylo zjistit, které vysvětlující proměnné jsou pro zjištění ceny domu přínosné, a které naopak nikoliv. Závěry z této fáze úkolu již nejsou vyvozeny na základě testů sloužících k posouzení statistické významnosti odhadnutých parametrů, jak tomu bylo v předchozí fázi. Nyní bude zkoumána významnost vysvětlujících proměnných z hlediska rozkladu regresního součtu čtverců. K této analýze slouží opět funkce `anova`, nyní ovšem pouze s jedním modelem jakožto argumentem. Pohled je zaměřen zejména na sloupec se sumami čtverců, platí, že čím vyšší regresní součet čtverců u dané vysvětlující proměnné, tím vyšší je její přínos v modelu.<sup>1</sup> Je ovšem známo, že příspěvek proměnných záleží na jejich pořadí v modelu, jelikož funkce `anova` provádí postupný rozklad na základě tohoto pořadí. V této fázi bylo proto vyzkoušeno více možností postupného rozkladu, viz výstupy pod textem.

```
#faze 6
#prvni zpusob:
anova(model_vlastni)

## Analysis of Variance Table
##
## Response: price
##
##              Df      Sum Sq   Mean Sq  F value
## poly(landValue, degree = 2, raw = TRUE)    2 4.7107e+12 2.3554e+12 596.5300
## livingArea                                1 3.8220e+12 3.8220e+12 967.9833
## newConstruction                          1 9.9810e+10 9.9810e+10  25.2781
## bedrooms                                1 6.9991e+10 6.9991e+10  17.7262
## livingArea:bedrooms                      1 1.1820e+10 1.1820e+10   2.9936
## Residuals                             1375 5.4291e+12 3.9485e+09
##
##              Pr(>F)
## poly(landValue, degree = 2, raw = TRUE) < 2.2e-16 ***
## livingArea                             < 2.2e-16 ***
## newConstruction                         5.617e-07 ***
## bedrooms                               2.717e-05 ***
## livingArea:bedrooms                     0.08382 .
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

---

<sup>1</sup>Poznámka: Součet sum čtverců u všech vysvětlujících proměnných dohromady dává regresní součet čtverců modelu. Dohromady s reziduální sumou čtverců pak získáváme celkový součet čtverců. Souvislost mezi indexem determinace a postupným rozkladem čtverců je proto zřejmý.

*#jiny zpusob rozkladu:*

```
anova(lm(price~ newConstruction + bedrooms + livingArea + livingArea:bedrooms
        + poly(landValue, degree = 2, raw = TRUE), data = train))
```

## Analysis of Variance Table

##

## Response: price

	Df	Sum Sq	Mean Sq	F value
## newConstruction	1	3.1958e+11	3.1958e+11	80.9374
## bedrooms	1	1.9036e+12	1.9036e+12	482.1202
## livingArea	1	5.2128e+12	5.2128e+12	1320.2069
## poly(landValue, degree = 2, raw = TRUE)	2	1.2666e+12	6.3330e+11	160.3915
## bedrooms:livingArea	1	1.1820e+10	1.1820e+10	2.9936
## Residuals	1375	5.4291e+12	3.9485e+09	
##		Pr(>F)		
## newConstruction		< 2e-16 ***		
## bedrooms		< 2e-16 ***		
## livingArea		< 2e-16 ***		
## poly(landValue, degree = 2, raw = TRUE)		< 2e-16 ***		
## bedrooms:livingArea		0.08382 .		
## Residuals				
## ---				
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

*#dalsi moznost rozkladu:*

```
anova(lm(price~ newConstruction + livingArea + bedrooms + livingArea:bedrooms
        + poly(landValue, degree = 2, raw = TRUE), data = train))
```

## Analysis of Variance Table

##

## Response: price

	Df	Sum Sq	Mean Sq	F value
## newConstruction	1	3.1958e+11	3.1958e+11	80.9374
## livingArea	1	6.9645e+12	6.9645e+12	1763.8450
## bedrooms	1	1.5195e+11	1.5195e+11	38.4821
## poly(landValue, degree = 2, raw = TRUE)	2	1.2666e+12	6.3330e+11	160.3915
## livingArea:bedrooms	1	1.1820e+10	1.1820e+10	2.9936
## Residuals	1375	5.4291e+12	3.9485e+09	
##		Pr(>F)		
## newConstruction		< 2.2e-16 ***		
## livingArea		< 2.2e-16 ***		
## bedrooms		7.297e-10 ***		
## poly(landValue, degree = 2, raw = TRUE)		< 2.2e-16 ***		
## livingArea:bedrooms		0.08382 .		
## Residuals				
## ---				
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Všechny tři vyzkoušené kombinace pořadí vysvětlujících proměnných se ovšem vcelku shodují – nejmenší součet čtverců se nachází ve všech případech u interakce, v tomto směru se tedy

tato fáze shoduje s předchozí. Dalšími proměnnými, které se z tohoto pohledu zdají být nejpostradatelnějšími, jsou pak proměnné `newConstruction` a `bedrooms`. U poslední ze zmíněných proměnných ale závisí na pořadí, ve kterém je model sestaven. Je možné pozorovat, že pokud se tato proměnná nachází v regresním modelu mezi prvními, tak její vliv značně stoupá. Možným vysvětlením této skutečnosti je silná korelace proměnné `bedrooms` s jinými vysvětlujícími proměnnými v modelu.

Na závěr je nutné dodat, že především v proměnné `newConstruction` nastává největší rozpor mezi výsledky z Fáze 5 a Fáze 6.

## Fáze 7

Úkolem poslední fáze této práce bylo nalézt takový model, který bude co nejpřesněji předpovídat ceny domu pro nová pozorování. Bylo vyzkoušeno mnoho různých modelů pomocí různých postupů (především s využitím backward stepwise regrese), ovšem žádný z modelů nesplňoval předpoklad homoskedasticity reziduí. Bylo proto nutné přistoupit k řešení tohoto problému, jelikož bez jeho eliminace nemohl vzniknout funkční model s dobrou predikční schopností. Z tohoto důvodu následně došlo k logaritmicizaci vysvětlované proměnné `price`, což vedlo k výraznému zlepšení.

Na základě předchozích analýz bylo odvozeno, že vhodnými vysvětlujícími proměnnými by mohly být například proměnné `waterfront`, `livingArea`, `landValue`, naopak například přidání proměnné `bedrooms` je na pováženou vzhledem k vysoké korelaci s proměnnou `livingArea`.

Příklad modelu s heteroskedastickými rezidui:

```
model_all23 = lm(price ~ (landValue + livingArea + waterfront + bedrooms) + landValue:livingArea ,
summary(model_all23) #ANO - r2 vychází lépe než u 2. mocniny
```

```
##
## Call:
## lm(formula = price ~ (landValue + livingArea + waterfront + bedrooms) +
##     landValue:livingArea, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -261167  -37628   -5909   29968  448954
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.400e+05  1.864e+04   7.513 1.04e-13 ***
## landValue      1.433e+00  1.403e-01  10.214 < 2e-16 ***
## livingArea     1.176e+02  4.827e+00  24.357 < 2e-16 ***
## waterfrontNo  -1.293e+05  1.750e+04  -7.386 2.62e-13 ***
## bedrooms      -1.147e+04  2.712e+03  -4.230 2.50e-05 ***
## landValue:livingArea -2.618e-04  5.738e-05  -4.562 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61940 on 1376 degrees of freedom
## Multiple R-squared:  0.6267, Adjusted R-squared:  0.6254
## F-statistic: 462.1 on 5 and 1376 DF, p-value: < 2.2e-16
```

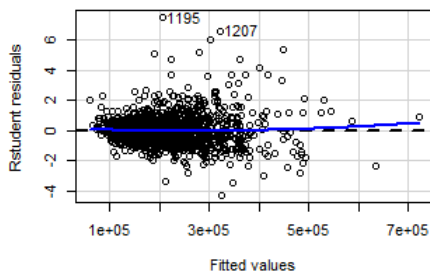
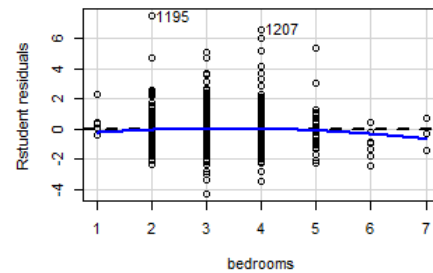
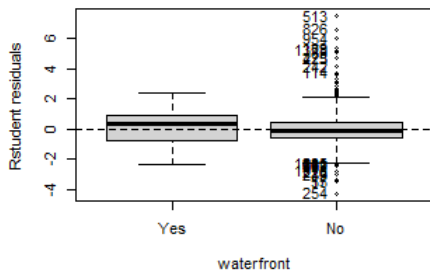
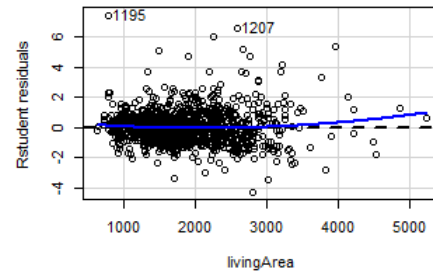
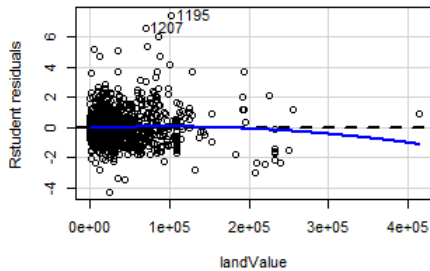


```
residualPlots(model_all123, tests = TRUE, id = TRUE, type = "rstudent")
```

```
##          Test stat Pr(>|Test stat|)
## landValue    -1.8137      0.069949 .
## livingArea     2.7612      0.005836 **
## waterfront
## bedrooms     -2.1349      0.032948 *
## Tukey test     1.8180      0.069060 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*#grafy o malinko horši nez f8*

*#testy: tukey sice nezam H0, ale p.h. jen 0,069 -> skoro rika, ze bychom meli neco pridať*



```
gqtest(model_all123, point = 0.5, fraction = 0.33, alternative = "greater",
       order.by = ~ livingArea, data = train) #zam H0: konst rezidua
```

```
##
## Goldfeld-Quandt test
##
## data: model_all123
## GQ = 2.976, df1 = 457, df2 = 456, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

```
# 3) Glejserův test
fit.d = lm(abs(resid(model_all23)) ~ (landValue + livingArea + waterfront + bedrooms) + landValue:livingArea, data = train)
#abs. hodnota rezidui se mení -> heterosk.
summary(fit.d) #zam H0: konst rezidua

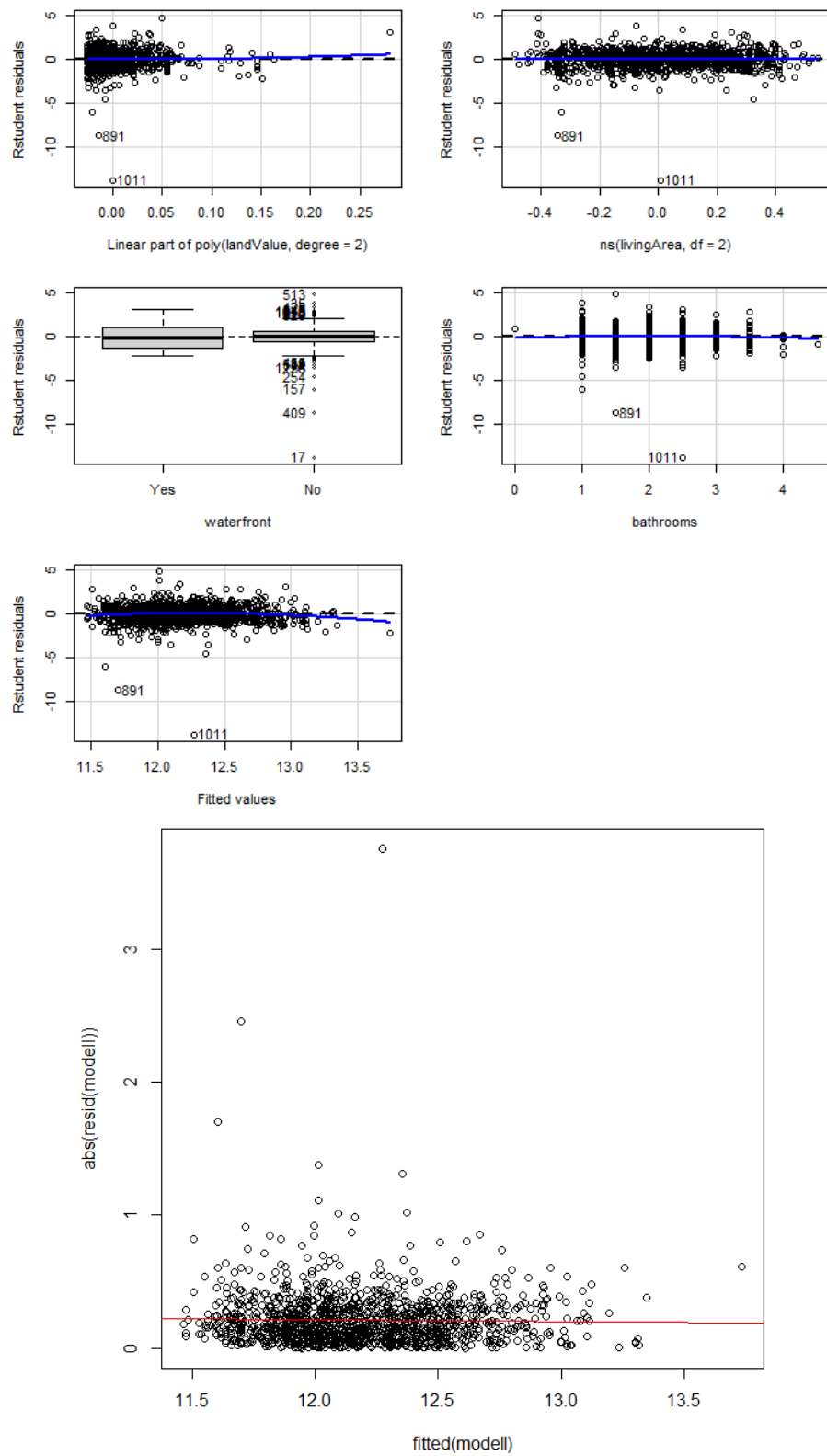
##
## Call:
## lm(formula = abs(resid(model_all23)) ~ (landValue + livingArea + waterfront + bedrooms) + landValue:livingArea, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -89559 -24011  -6484   14851 393283
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.113e+04  1.224e+04   1.727  0.08438 .
## landValue      4.634e-01  9.211e-02   5.031 5.52e-07 ***
## livingArea     2.765e+01  3.169e+00   8.725 < 2e-16 ***
## waterfrontNo  -1.014e+04  1.149e+04  -0.882  0.37774
## bedrooms      -7.460e+03  1.781e+03  -4.189 2.98e-05 ***
## landValue:livingArea -1.207e-04  3.768e-05  -3.204  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40670 on 1376 degrees of freedom
## Multiple R-squared:  0.1268, Adjusted R-squared:  0.1236
## F-statistic: 39.95 on 5 and 1376 DF,  p-value: < 2.2e-16
```

Příklad modelu s homoskedastickými rezidui:

```
## Call:
## lm(formula = log(price) ~ poly(landValue, degree = 2) + ns(livingArea, df = 2) + waterfront + bathrooms, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.7554 -0.1619  0.0114  0.1718  1.3804
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.01450    0.08841 135.902 < 2e-16 ***
## poly(landValue, degree = 2)1    4.55639    0.32689  13.939 < 2e-16 ***
## poly(landValue, degree = 2)2   -2.00275    0.30066   -6.661 3.92e-11 ***
## ns(livingArea, df = 2)1         1.56250    0.09113  17.145 < 2e-16 ***
## ns(livingArea, df = 2)2         0.70662    0.12168   5.807 7.87e-09 ***
## waterfrontNo    -0.55522    0.08372  -6.632 4.74e-11 ***
## bathrooms       0.11771    0.01747   6.736 2.39e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2915 on 1375 degrees of freedom
## Multiple R-squared:  0.5905, Adjusted R-squared:  0.5887
## F-statistic: 330.5 on 6 and 1375 DF,  p-value: < 2.2e-16
residualPlots(model_log8, tests = TRUE, id = TRUE, type = "rstudent")

## Warning in residualPlot.default(model, ...): Splines replaced by a fitted linear
## combination

##              Test stat Pr(>|Test stat|)
## poly(landValue, degree = 2)
## ns(livingArea, df = 2)
## waterfront
## bathrooms      -0.9875      0.3235
## Tukey test      -4.8993    9.617e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```

# ad d) Goldfeld-Quandt test - vse ok :)
gqtest(model_log8, point = 0.5, fraction = 0.33, alternative = "greater",
       order.by = ~ landValue, data = train)

##
## Goldfeld-Quandt test
##
## data: model_log8
## GQ = 1.1189, df1 = 456, df2 = 455, p-value = 0.1156
## alternative hypothesis: variance increases from segment 1 to 2
gqtest(model_log8, point = 0.5, fraction = 0.33, alternative = "greater",
       order.by = ~ livingArea, data = train)

##
## Goldfeld-Quandt test
##
## data: model_log8
## GQ = 0.72314, df1 = 456, df2 = 455, p-value = 0.9997
## alternative hypothesis: variance increases from segment 1 to 2
gqtest(model_log8, point = 0.5, fraction = 0.33, alternative = "greater",
       order.by = ~ waterfront, data = train)

##
## Goldfeld-Quandt test
##
## data: model_log8
## GQ = 0.51386, df1 = 456, df2 = 455, p-value = 1
## alternative hypothesis: variance increases from segment 1 to 2
gqtest(model_log8, point = 0.5, fraction = 0.33, alternative = "greater",
       order.by = ~ bathrooms, data = train)

##
## Goldfeld-Quandt test
##
## data: model_log8
## GQ = 0.65847, df1 = 456, df2 = 455, p-value = 1
## alternative hypothesis: variance increases from segment 1 to 2

# ad e) Glejserův test - ok

fit.d = lm(abs(resid(model_log8)) ~ poly(landValue, degree=2)+ ns(livingArea, df=2)+w:
summary(fit.d)

##
## Call:
## lm(formula = abs(resid(model_log8)) ~ poly(landValue, degree = 2) +
##     ns(livingArea, df = 2) + waterfront + bathrooms, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.2782 -0.1227 -0.0411  0.0771  3.5560
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.330710   0.061593   5.369 9.27e-08 ***
## poly(landValue, degree = 2)1  0.122103   0.227744   0.536  0.592
## poly(landValue, degree = 2)2  0.307707   0.209472   1.469  0.142
## ns(livingArea, df = 2)1    -0.078746   0.063493  -1.240  0.215
## ns(livingArea, df = 2)2     0.099879   0.084772   1.178  0.239
## waterfrontNo    -0.086691   0.058326  -1.486  0.137
## bathrooms        0.004017   0.012175   0.330  0.741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2031 on 1375 degrees of freedom
## Multiple R-squared:  0.00933, Adjusted R-squared:  0.005007
## F-statistic: 2.158 on 6 and 1375 DF, p-value: 0.04453

```

```

BIC(model_log9)

## [1] 638.8412
BIC(model_log8)

## [1] 565.7288
BIC(model_log7)

## [1] 601.2504
#BIC NEJNIZSI PRO m8 ->m7 ->m9

```

Po transformaci vysvětlované proměnné následoval nový výběr vhodného modelu, opět bylo nahlíženo především na index determinace (vzhledem k různému počtu parametrů to byl především adjustovaný koeficient determinace), informační kritéria (vzhledem k různému počtu parametrů to bylo především BIC) a počet parametrů modelu a hledala se ideální kombinace všech těchto hodnot. Indexy determinace se pohybovaly kolem 60 %, BIC kolem 600. Velice vhodným modelem se zdál být model s názvem model\_log8, který měl nízké BIC, vcelku vysoký adjustovaný index determinace (58,87 %, pouze na základě většího počtu parametrů by se o něm mohlo pochybovat v porovnání s jinými modely. Tento model byl dále testován na heteroskedasticitu, přičemž i těmito testy model úspěšně prošel.

Dále byla zkoumána odlehlá a vlivná pozorování v tomto modelu, které poukázalo na několik odlehlých pozorování, ovšem tato pozorování nebyla určena jako vlivná a proto nebylo nutné se jimi dále zabývat. Jako vlivné se na základě Cookovy vzdálenosti větší než 1 ukázalo pouze jedno pozorování s číslem 702, jehož Cookova vzdálenost přesahovala hodnotu 1 i přes to, že hodnota vnějšně studentizovaného rezidua byla velmi nízká, hlavní vliv na vysokou Cookovu vzdálenost měla proto vysoká hodnota  $h_{ii}$ , jak ukazuje také diagnostický graf níže. Bylo vyskoušeno, jak moc ovlivní vynechání tohoto pozorování výsledky odhadů regresních koeficientů – změny byly ovšem jen velmi malé.

```

#ODLEHLA POZOROVANI - DIVAM SE NA P-H. U BONF KOREKCE!
# H0: v souboru není odlehlé pozorování z hlediska regresního modelu
outlierTest(model_log8)

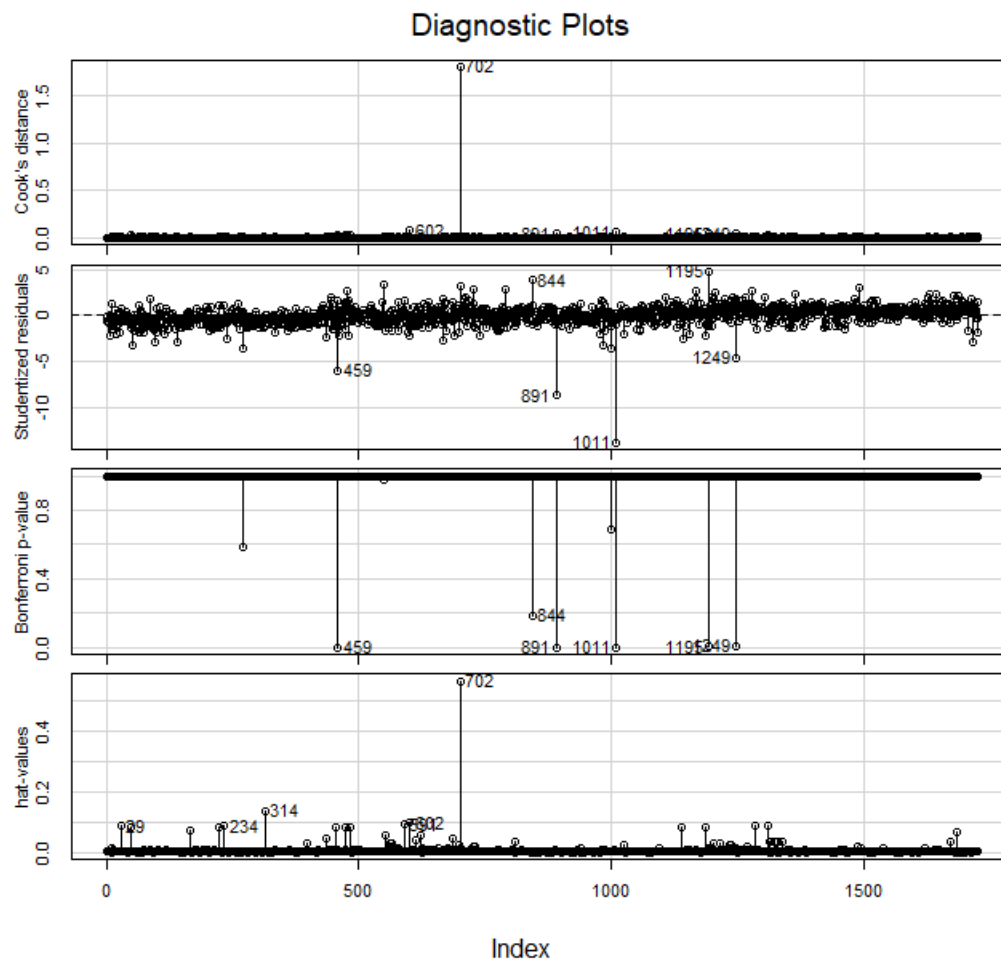
##          rstudent unadjusted p-value Bonferroni p
## 1011 -13.751830      2.0643e-40    2.8528e-37
## 891  -8.671265      1.1841e-17    1.6364e-14
## 459  -5.924277      3.9592e-09    5.4716e-06
## 1195  4.803726      1.7281e-06    2.3882e-03
## 1249 -4.564697      5.4492e-06    7.5308e-03

# Vlivná pozorování: Cookova vzdálenost

sort(cooks.distance(model_log8), decreasing = TRUE) # nejv. má D=1,818 = C.702

##          702          602          1011          1249          1195          891
## 1.817512e+00 7.498079e-02 5.818392e-02 4.246657e-02 4.173666e-02 4.086827e-02

```



Posledním krokem před samotou predikcí bylo prověření nemultikolinearity pomocí indexu VIF - jeho hodnoty vyšly o dost menší než 5, tudíž nebyla v modelu prokázána multikolinearita.

```
#MULTIKOLINEARITA: OK :)
vif(model_log8)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## poly(landValue, degree = 2) 1.336858 2      1.075279
## ns(livingArea, df = 2)      2.490246 2      1.256205
## waterfront                  1.062013 1      1.030540
## bathrooms                   2.165951 1      1.471717
```

V závěru úlohy byla s využitím finálního modelu `model_log8` na testovacích datech odhadnuta očekávaná čtvercová chyba předpovědi ceny nemovitosti (ve výstupu označena jako MSE).

```

# Testovací datový soubor
chyby = log(test$price) - predict(model_log8, test) # chyby předpovědi
hist(chyby)
abline(v=0, col="red")

plot(predict(model_log8, test) ~ log(test$price), xlab = "skutečnost",
      ylab = "predikce", col = "blue", cex = 1.5); abline(a = 0, b = 1,
      col = "red")

(ME = mean(chyby))

## [1] -0.02661301
(MSE = mean(chyby^2))

## [1] 0.09113131
sqrt(MSE)

## [1] 0.3018796

```