Zápočtová úloha z 01REAN, 19. 12. 2018

Pro získání zápočtu z předmětu 01REAN je potřeba vypracovat zápočtovou úlohu, tj. odevzdat protokol ve formátu `pdf` a příslušný *R Markdown* soubor ve formátu `Rmd`, kde budou zodpovězeny **všechny** otázky. Veškeré otázky se týkají datového souboru *Car_data.csv*, popisující odhadní prodejní cenu ojetých méně než rok starých aut z koncernu GM v roce 2005.

Popis datového souboru Car data

Datový soubor `Car_data.csv` obsahuje 805 pozorování o 12 proměnných. Autorem souboru je Shonda Kuiper a data byly posbírány na použitých autech v roce 2005, kde jim byla na základě *Central Edition of the Kelly Blue Book* přiřazena prodejní cena.

Anglický popis jednotlivých proměnných (sloupců) je následující:

Feature	Description
`Price`	suggested retail price of the used 2005 GM car in excellent condition.
`Mileage`	number of miles the car has been driven.
`Make`	manufacturer of the car such as Saturn, Pontiac, and Chevrolet.
`Model`	specific models for each car manufacturer such as Ion, Vibe, Cavalier.
`Trim`	specific type of car model such as SE Sedan 4D, Quad Coupe 2D.
`Type`	body type such as sedan, coupe, etc.
`Cylinder`	number of cylinders in the engine.
`Liter`	measure of engine size.
`Doors`	number of doors.
`Cruise`	indicator variable representing whether the car has cruise control. 1 = presence of cruise control.
`Sound`	indicator variable representing whether the car has upgraded speakers. 1 = presence of upgraded speakers.
`Leather`	indicator variable representing whether the car has leather seats. 1 = presence of leather seats.

Během práce budete potřebovat řadu balíků (knihoven), které jsme během semestru využívali. Pro načtení a nainstalování potřebných balíků můžete použít tento kód:

```
load.libraries <- c('car','MASS','ggplot2','ISLR','graphics','effects','lattice',
'leaps','psych','lmtest','robustbase')
install.lib <- load.libraries[!load.libraries %in% installed.packages()]
for(libs in install.lib) install.packages(libs, dependences = TRUE)
sapply(load.libraries, require, character = TRUE)
```

Požadavky k vypracování a odevzdání

Zápočtovou úlohu vypracujte a odevzdejte samostatně. V případě konzultace a spolupráce s kolegy, uveďte u dané otázky s kým vším jste na daném řešení spolupracovali. Pokud obdržím shodný kód se stejným vysvětlením postupu a nebude uveden spolupracující (původní) autor, tak budu považovat danou otázku za nezodpovězenou. Vypracované řešení (tj. `pdf` soubor a příslušný `Rmd` soubor) odešlete e-mailem na adresu jiri.franc@fjfi.cvut.cz.
V protokolu každou otázku otázku označte jako novou sekci, zkopírujte text zadání a poté vypracujte odpověď. Všechny odpovědi včetně numerických hodnot musí být uvedeny v textu, ale v protokolu musí být jasné odkud dané hodnoty pochází (výstupy ze summary funkcí, obrázky, výstupy z testů hypotéz apod.).

Odevzdání protokolu s odpovědmi na všechny otázky a s tím související získání zápočtu je podmínkou pro zapsání zkoušky do KOSu.

Zadání

Vypracujte následující body zadání a zodpovězte příslušné otázky. Hlavním cílem je vytvořit lineární model pro predikci ceny ojetého automobilu na základě dostupných proměnných:

Průzkumová a grafická část:

Q01: Zjistěte, zdali data neobsahují chybějící hodnoty (`NA`), pokud ano tak rozhodněte zdali můžete příslušná pozorování z dat odstranit. Které proměnné jsou kvantitativní a které kvalitativní? Jeli možno některé zařadit do obou skupin, pro kterou se rozhodnete? Které proměnné budete brát jako faktorové a proč?

Q02: Proměnnou `Mileage` nahraďte proměnnou `Odometer` kde bude uveden stav tachometru v kilometrech místo v mílích. Vysv2tlovanou proměnnou `Price` převeďte z USD na CZK. Vykreslete histogramy a odhady hustot pro tyto dvě proměnné, tj `Price` a `Odometer`.

Q03: Pro proměnné `Price, Odometer` vykreslete `scatterplot` - závislost odezvy na vysvětlující proměnné a proložte body jak lineárním odhadem tak vyhlazenou křivkou lokální regrese, buď pomocí LOESS (locally

estimated scatterplot smoothing) nebo LOWESS (locally weighted scatterplot smoothing) (`lines(lowess(X,Y))`). Co lze z tohoto obrázku tvrdit o závislosti ceny auta na počtu najetých km?

Q04: Pro proměnné `Make, Type, Doors, Cruise, Sound, Leather` a jejich vztah k odezvě `Price` vykreslete krabicové diagramy (boxploty). Je mezi uvedenými proměnnými některá, pro kterou byste na základě krabicových diagramů navrhli sloučit určité úrovně dohromady?

Q05: Pro kombinaci faktorizovaných proměnných `Make` a `Model` vykreslete cenu aut, aby bylo na obrázku vidět, jestli se liší ceny u jednotlivých modelů v závislosti na výrobci a naopak jak se liší ceny u jednotlivých výrobců v závislosti na modelu.

Q06: Pro auta výrobce SAAB vykreslete závislost ceny na počtu ujetých kilometrů, kde jednotlivé události označíte barvou podle Typu auta a velikost bodů v grafu bude odpovídat objemu motoru.

Q07: Navrhněte další zobrazení datového souboru. Proveďte ho a popište jeho účel.

Regresní model závislosti ceny automobilu Saturn na počtu najetých km.

Q08: Sestavte jednoduchý regresní model a na jeho základech zjistěte zdali cena ojetého automobilu značky Saturn závisí na počtu najetých kilometrů. Pokud ano, o kolik se změní odhadovaná cena automobilu Saturn při najetí 1000km navíc? Ověřte předpoklady pro použití lineárního modelu a diskutujte výstup.

Q09: Dá se předešlý jednoduchý regresní model zlepšit pomocí logaritmické transformace odezvy? Jak se poté změní (navýší/poklesne) cena automobilů při změně počtu najetých kilometrů o 1000 km? Zdůvodněte proč případná transformace je přínosná, nebo naopak nepřínosná.

Q10: Vyzkoušejte transformovat nezávislou proměnnou `Odometer`. Vyzkoušejte například po částech konstantní transformaci, splines a polynomiální transformaci (kvadratickou a kubickou).

Q11: Vykreslete scatterplot skutečných cen aut a stavu tachometru a na základě vybraného modelu, proložte skrze data odhadnutou regresní přímku a vykreslete 95% konfidenční intervaly jak pro predikované hodnoty, tak pro re-

gresní přímku (tzv. Confidence a Prediction band). Porovnejte s výsledkem z funkce `plot(allEffects(model))`.

Q12: Přidejte k vysvětlujícím proměnným i `Type` a `Model`, navrhněte aditivní lineární model, a ve scatterplotu vykreslete jednotlivé skupiny různými barvami a data proložte odpovídajícími regresními přímkami.

Q13: Proveďte validaci modelu z bodu 11 pomocí příslušných testů na rezidua a pomocí příslušných obrázků (QQplot, residua vs. fitted, atd.)

Vícerozměrný regresní model:

Q14: Porovnejte pomocí vhodného statistického testu shodnost středních hodnot cen u automobilů mající a nemající tempomat, u atomobilů mající a nemající vylepšené reproduktory a mající a nemající sedačky v kůži. Zdůvodněte, zdali tyto statistické testy jsou vypovídající a zdali lze z nich určit důležitost daných proměnných pro predikci ceny automobilu.

Q15: Zkonstruujte lineární model popisující cenu automobilů GM s využitím všech dostupných proměnných, kde bude přítomna interakce nejvýše dvou proměnných. Na základě kritérií jako jsou `AIC, BIC,` R^2`, F, atd.` vyberte nejvhodnější model. Ten validujte a okomentujte jeho výběr.

Q16: Pro vybraný model z bodu 15, vyzkoušejte jak logaritmickou transformaci odezvy `Price`, tak Box-Coxovu transformaci pro zlepšení normality reziduí. Vykreslete optimální log-věrohodnostní profil u Box-Coxovy transformace a porovnejte navrženou mocninou transformaci s logaritmickou. Pro který model se rozhodnete a proč?

Q17: Pro model s log transformovanou cenou v bodě 15 vyčtěte procentuální navýšení/pokles ceny automobilů při změně počtu najetých kilometrů o 1000 km. Porovnejte jak se změnil vliv počtu najetých kilometrů na cenu automobilů Saturn v porovnání s modelem z bodu 9.

Q18: Spočtěte korelace mezi všemi dostupnými proměnnými, kde to dává smysl a u korelovaných proměnných se pokuste zdůvodnit důvod této korelace. Zkoumejte případnou multikolinearitu ve vašem finálním modelu z bodu 16 a pomocí podmíněnosti matice regresorů, VIF a dalších nástrojů validujte váš výběr.

Q19: Vykreslete tzv. *Partial regression plot* a tzv. *Partial residual plot* pro finální model, okomentujte co nám zmíněné grafy říkají o výsledném modelu.

Q20: Prezentujte váš výsledný model pro predikci `Price`, diskutujte výsledné parametry R^2 a σ tohoto modelu. Validujte model (jak graficky, tak pomocí příslušných testů hypotéz).

Robustní regrese

Q21: Obsahuje Váš model z bodu 19 nějaká vlivná pozorování? Pokuste se detekovat odlehlé a pákové body pomocí různých diagnostických nástrojů tzv. *leave-one-out deletion regression*.

Q22: Pokud jste odhalili nějaká vlivná pozorování, jak byste s nimi naložili a proč?

Q23: Porovnejte regresní koeficenty, které jste obdrželi z výsledného klasického lineární modelu (s a bez odlehlých pozorováních) s robustními modely. Vyzkoušejte MM odhad (pro dva druhy funkce ψ) a LTS odhad (při použití 90% a 50% pozorování).

Kam dál?

Q24: Diskutujte jak by šlo případně zlepšit predikci, jaké transformace jednotlivých proměnných by mohli pomoci. Převedli byste některé další spojité proměnné na diskrétní (na faktory)? Jaké další kroky byste při analýze navrhli?

Q25: Představte si, že jste členem mafie, která provozuje síť autobazarů. Od bosse dostanete úkol spočítat, zdali se vyplatí stáčet u aut odometry. Bazary provozujete ve státě s pochybným právním prostředím, kde stočení tachometru není trestný čin a tudíž hrozí jen pokuta. Pokud vaše prodeje odpovídají zkoumaným datům, jak vysoká by musela být pokuta, aby se nevyplatilo stáčet odometr? Předpokládejte, že Vás v průměru odhalí u každého 500 auta a v průměru budete stáčet odometry o 10000 km.