1 VAJE: Več številskih in opisnih napovednih spremenljivk

1.1 `Carseats`

V podatkovnem okviru `Carseats` v paketu `ISLR` so podatki za 400 trgovskih centrov v ZDA, ki prodajajo tudi avtomobilske otroške stolčke. Upoštevali bomo naslednje spremenljivke, podatki so za določeno koledarsko leto:

- `Sales` predstavlja letno število prodanih stolčkov (v 1000) za posamezni trgovski center;
- `Price` (USD) je cena posameznega stolčka v trgovskem centru;
- `Advertising` predstavlja letne stroške oglaševanja (1000 USD) za posamezni trgovski center;
- `ShelveLoc`, je kakovost police, na kateri prodajajo otroške stolčke, njene vrednosti so: Bad, Medium, Good.

Zanima nas odvisnost `Sales` od `Price`, `Advertising` in `ShelveLoc`. Ali je pri vseh vrednostih `ShelveLoc` odvisnost od `Price` in `Advertising` enaka?

1. Grafično prikažite odvisnost `Sales` od `Price`, `Advertising` in `ShelveLoc` in grafikone na kratko obrazložite.
2. Ali grafični prikaz narekuje obstoj interakcije med `ShelveLoc` in `Price` in/ali obstoj interakcije med `ShelveLoc` in `Advertising`? Kako se to vidi?
3. Zapišite model, za odvisnost `Sales` od `Price`, `Advertising` in `ShelveLoc` brez interakcij ter model, v katerega dodamo tudi dve interakciji `ShelveLoc:Price` in `ShelveLoc:Advertising`. Kaj geometrijsko predstavljata ta dva modela? Ali za modela lahko rečemo, da sta gnezdena?
4. Uporabite F-test za primerjavo dveh gnezdenih modelov, da ugotovite, ali interakciji doprineseta k izboljšanju modela. Zapišite ničelno domnevo in obrazložite rezultat testiranja.
5. Nadaljujemo z modelom, ki ga narekuje F-test, ki ste ga izvedli v prejšnji točki. Ali bi bilo potrebno v model vključiti interakcijski člen `Price:Advertising`? Kateri grafični prikaz omogoča diagnostiko interakcije dveh številskih spremenljivk? Uporabite ta grafični prikaz in ga na kratko obrazložite. Utemeljite svojo odločitev s statističnim testom.
6. Naredite diagnostiko v prejšnji točki izbranega modela.
7. Pokažite, kaj predstavlja graf dodane spremenljivke, če je ta opisna.

8. Preverite statistično značilnost spremenljivke `ShelveLoc` ob upoštevanju ostalih dveh spremenljivk v modelu. Zapišite ničelno domnevo, ki jo testirate.

9. Hkratno testirajte osnovne ničelne domneve v modelu in obrazložite rezultate.

10. Napišite matriko primerjav, s katero hkrati testirate vse pare razlik položajev polic (`ShelveLoc`) in statistično značilnost `Price` in `Advertising`.

1.2 `pacienti`

v datoteki `PACIENTI1.txt` so podatki za 20 pacientov s povišanim krvnim tlakom. Za vsakega pacienta so navedene vrednosti naslednjih spremenljivk: zgornji krvni tlak (`SKT`, mm Hg), starost (`starost`, leta), telesna masa (`masa`, kg), površina telesa (`PT`, m^2), bazalni srčni utrip (`utrip`, število utripov na minuto) in stresni indeks (`stres`). Zanima nas odvisnost `SKT` od vseh ostalih spremenljivk.

1. Analizirajte povezanost vseh spremenljivk v podatkovnem okviru `pacienti`. V ta namen podatke ustrezno grafično predstavite in izračunajte koeficiente korelacije. Obrazložite rezultate.

2. Naredite model za odvisnost SKT od vseh ostalih spremenljivk in preverite, ali je v modelu prisotna kolinearnost. Kakšne probleme lahko povzroča kolinearnost v modelu?

3. Izberite spremenljivko, ki jo je zaradi kolinearnosti potrebno izločiti iz modela. Pri tem upoštevajte, da je `PT` izračunana na podlagi `masa` in telesne višine, ki je sicer med podatki ni, po formuli:

   $PT = 0.007184 \cdot visina^{0.725} \cdot masa^{0.425}$.

4. Primerjajte ocene parametrov modela, kjer je prisotna kolinearnost in modela brez kolinearnosti. Ali so razlike v skladu s pričakovanji?

5. Obrazložite končni model.

1.3 Spanje

V datoteki SLEEP.txt (manjkajoči podatki označeni z NA) so podatki za 62 sesalcev. Glej `http://www.statsci.org/data/general/sleep.html`. Delno to informacijo poznamo iz podatkovnega okvira `mammals`. Zanima nas, kako je `TotalSleep` (h/dan) odvisen od `logBodyWt` (kg), `log(BrainWt)` (g), `Gestation` (dnevi), `LifeSpan` (leta) in `Danger3`. Pričakujemo, da med napovednimi spremenljivkami obstaja povezanost.

Analizirajte odvisnost `TotalSleep` od `logBodyWt`, `log(BrainWt)`, `Gestation`, `LifeSpan` in `Danger3`.

Katere spremenljivke je smiselno vključiti v model, da ne bo težav s kolinearnostjo?

Ali je potrebna transformacija odzivne ali napovednih spremenljivk?

Ali bi bilo smiselno v model vključiti kakšno interakcijo?

Prikažite korake modeliranja in obrazložite izbrani model.

1.4 Poraba goriva na avtocestah

Raziskovalno vprašanje: kako je poraba goriva na avtocestah odvisna od lastnosti avtomobila?

Raziskovalne domneve: poraba goriva na avtocestah je odvisna od

- tehničnih karakteristik avta (masa, prostornina, moč): večji avti imajo večjo porabo;
- od tipa avta: večji avti imajo večji upor in s tem večjo porabo;
- od porekla avta: avti iz ZDA imajo večjo porabo kor ne-ZDA avti.

Podatki: v paketu MASS je datoteka `Cars93` s karakteristikami avtomobilov, glej `help(Cars93)`.

```
> library(MASS)
> # help(Cars93) ## Data from 93 Cars on Sale in the USA in 1993
> # names(Cars93)
```

Izbrane spremenljivke `MPG.highway, Weight, EngineSize, Horsepower, Type` in `Origin` spremenimo v nam razumljive merske enote in uporabimo slovenska imena spremenljivk.

```
> Cars93$Poraba<-235.21/Cars93$MPG.highway  # v l/100 km
> Cars93$Masa<-Cars93$Weight*0.45359/100    # v 100 kg
> Cars93$Prostornina<-Cars93$EngineSize     # v litih
```

```
> Cars93$Moc<-Cars93$Horsepower              # v KM
> Cars93$Poreklo<-Cars93$Origin
> Cars93$Tip<-Cars93$Type
```

Naredimo nov podatkovni okvir `avti` z izbranimi spremenljivkami.

```
> avti <- subset(Cars93, select=c(Poraba, Masa, Prostornina, Moc, Poreklo, Tip))
> rownames(avti)<-Cars93$Make   ### identifikator vozila na slikah
```

Tip `Van` je v več pogledih drugačen od ostalih tipov avtov (večja površina in drugačne lastnosti motorja), vse ostale tipe avtov bi radi primerjali s tipom `Van`, zato ga vzamemo za referenčno skupino.

```
> avti$Tip<-relevel(avti$Tip, ref="Van")
```

Naredite ustrezni model in ga obrazložite.