Linearna regresija u aktuarstvu

Škiljan, Paula

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Paula Škiljan

LINEARNA REGRESIJA U AKTUARSTVU

Diplomski rad

Voditelj rada:
doc.dr.sc. Siniša Slijepčević

Zagreb, Rujan 2019.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Ovaj rad posvećujem dvjema najbitnijim osobama u svom životu: zaručniku Domagoju koji mi je bio najveća podrška tokom obrazovanja i sestri Ani bez koje je bilo što nezamislivo.

Sadržaj

Poglavlje 1

Uvod

Statistika je znanost o prikupljanju, sažimanju, obradi i analizi podataka u svrhu donošenja zaključaka o stvarnom svijetu. Ona omogućuje korištenje naprednih matematičkih tehnika u raznim drugim područjima i disciplinama kao što su aktuarska znanost ili financije te je stoga njezina rastuća važnost u današnjem svijetu neupitna. Zbog velikog stupnja informatizacije, prikupljanje velike količine podataka je postala svakodnevica te je umijeće obrade i analize podataka sve bitnije [1].

Podaci su informacije o okolini koje su transformirane na numeričku skalu, tzv. numerički podaci ili poprimaju jednu od konačno mnogo vrijednosti koje predstavljaju različite kategorije, tzv. kategorički podaci. Kako bismo opisali naše podatke i mogli donositi zaključke s pouzdanošću, podaci se moraju opisati pomoću matematičkog modela. Tek pomoću modela koji je oblikovan podacima mogu se donositi novi zaključci. Kompleksnost modela kojeg koristimo ovisi o našem predznanju i vjerovanju o tome što podaci predstavljaju. Model treba biti dovoljno kompleksan kako bi uspješno mogao opisati podatke. Međutim, veća jednostavnost modela nam omogućava veću dubinu razumijevanja matematičkih svojstava te veću pouzdanost u opravdanost korištenja danog modela. Također su jednostavniji modeli polazna točka za razne kompleksnije modele te ih je stoga važno dobro savladati.

1.1 Regresijska analiza

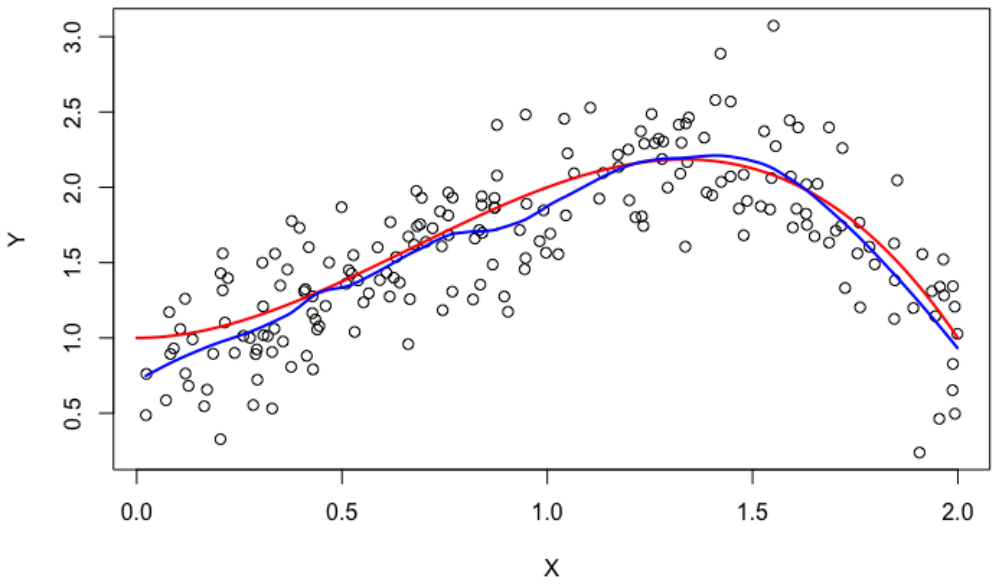
Regresijska analiza je jedna od najbitnijih grana statistike u kojoj je cilj saznati informacije o distribuciji neke istaknute varijable Y, koja se naziva odaziv, u ovisnosti o vrijednostima koje poprimaju druge, eksplanatorne varijable $X_1, \ldots, X_p$, koje se još nazivaju i prediktori. Glavna značajka regresijske analize je sposobnost davanja izjava o varijablama nakon provedenih kontroliranih promjena poznatih

eksplanatornih varijabli.

Ako pretpostavimo da $\text{Var}(Y|X_1 = x_1, \ldots, X_p = x_p)$ ne ovisi o vrijednostima prediktora $x_1, \ldots, x_p$, onda možemo zapisati

$$Y = f(X_1, \ldots, X_p) + \epsilon,$$

pri čemu je izraz $f(X_1, \ldots, X_p)$ očekivanje odaziva Y u ovisnosti o vrijednostima prediktora $X_1, \ldots, X_p$, a ϵ je slučajna varijabla s očekivanjem 0 koja opisuje grešku, tj. odstupanje od očekivanja. Normalna distribucija greške je imala ključnu ulogu u razvoju regresijske analize i najčešća je pretpostavka, ali od interesa je proširiti ideje regresije i na druge modele podataka.



Slika 1.1: Na slici vidimo ovisnost odzivne varijable Y o prediktoru X. Crvena krivulja predstavlja uvjetno očekivanje od Y u ovisnosti o X, a plava krivulja je regresijska krivulja određena na temelju danog uzorka lokalnom polinomnom regresijom.

1.2 Važnost i primjenjivost linearne regresije

Ako modeliramo odaziv kao $Y = f(X_1, \ldots, X_p) + \epsilon$, moramo i modelirati oblik funkcije f, tj. pretpostavljamo da je $f \in \mathcal{F}$, gdje je $\mathcal{F}$ neki skup funkcija. Ukoliko bi

f smio biti proizvoljan, došlo bi do preprilagodbe modela te bi regresijska funkcija prolazila savršeno kroz sve dane trening podatke, no greška bi bila velika na test podacima.

Jedan od najjednostavnijih modela $\mathcal{F}$ je linearni model gdje je očekivanje odaziva neka linearna kombinacija eksplanatornih varijabli:

$$f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Pri čemu se član β_0 naziva odsječak (eng. intercept) te se najčešće dodaje kako analiza ne bi ovisila o translaciji varijabli, npr. da li se temperatura mjeri u fahrenheitima ili u stupnjevima celzijusa.

Razlozi za primjenu linearne regresije su:

- Jednostavna prilagodba modela metodom najmanjih kvadrata.
- Jednostavna intepretacija veze odzivne i eksplanatornih varijabli. Ukoliko se promijeni vrijednost eksplanatorne varijable, lako je procijeniti kako će se promijeniti vrijednost odzivne.
- Duboko matematičko razumijevanje statističkih svojstava, što omogućuje
  - jednostavno računanje vrijednosti raznih procjenitelja i njihovih distribucija,
  - konstrukciju intervala pouzdanosti za procijenjene parametre modela, koji nam daju dojam o mogućoj pogrešci koja je nastala prilikom prilagodbe modela,
  - predviđanje budućih vrijednosti odaziva te konstrukciju predikcijskih intervala za njih.

Transformacija varijabli je bitna tehnika kojom je moguće prilagoditi mnogo kompleksnije modele. Moguće je prilagoditi bilo kakve modele koji su linearna kombinacija raznih transformacija eksplanatornih varijabli, npr.

$$f(X_1, X_2, X_3) = \beta_0 + \beta_1 X_1^3 + \beta_2 e^{X_2} + \beta_3 \sin(X_3) + \beta_4 X_1 \log(X_2)$$

Također, veoma bitan i primjenjiv primjer u praksi je polinomijalna regresija gdje odzivnu varijablu Y modeliramo polinomom stupnja p u ovisnosti o varijabli X pa je $X_p = X^p$:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p.$$

Prilikom analize odzivnih varijabli (npr. imovina firmi, plaće pojedinaca, vrijednost nekretnina) u primjenjenoj statistici uobičajeno je promatrati i logaritmirane vrijednosti. Naime, logaritmiranje održava originalni poredak, ali značajno umanjuje ekstremne vrijednosti distribucije. Npr. ako vjerujemo da svakom godinom radnog iskustva ili obrazovanja plaća u prosjeku raste oko 10%, onda možemo prilagoditi linearan model:

$$\log(\text{plaća}) = \beta_0 + \beta_1 \cdot \text{obrazovanje} + \beta_2 \cdot \text{iskustvo} + \epsilon$$

1.3 Regresija u aktuarstvu

Statistička analiza podataka je glavni alat u modernoj aktuarskoj znanosti [2]. Tri tradicionalna područja aktuarstva su određivanje premije, rezervacije (eng. reserving) i testiranje solventnosti. Regresijska analiza igra važnu ulogu u raznim područjima aktuarske znanosti, a linearni model je ishodišna točka za prilagodbu modela podacima.

Određivanje premije

Regresijska analiza se može koristiti za određivanje cijene raznih polica osiguranja. Npr. u slučaju osiguranja osobnog automobila premija ovisi o dobi, spolu, mjestu stanovanja, namjeni vozila (za posao ili osobne potrebe) i raznim drugim faktorima. U tom slučaju se regresija može koristiti kako bi se odredile varijable koje imaju bitnu ulogu u određivanju očekivanih potraživanja.

Na kompetitivnom tržištu osiguravajuće kuće ne koriste iste cijene za sve osiguranike. Naime, kada bi postojale jedinstvene cijene, niskorizični klijenti (s ispodprosječnim potraživanjima) bi preplaćivali svoje osiguranje te bi stoga odlučili promjeniti osiguravajuću kuću. S druge strane, visokorizični klijenti (klijenti s iznadprosječnim potraživanjima) bi ostali. Tada bi moralo doći do rasta premija zbog povećanog potraživanja od strane većeg udjela rizičnih kijenata. To dovodi do ljuljanja tržišnog udjela jer je tvrtka izgubila niskorizične klijente. Ovaj problem je poznat pod nazivom nepovoljni odabir (eng. adverse selection). Korištenjem prikladnog skupa eksplanatornih varijabli može se razviti klasifikacijski sustav tako da svaki osiguranik plaća svoj pravedan udio.

Rezerviranje i testiranje solventnosti

Rezerviranje podrazumijeva određivanje prikladne količine kapitala za podmirivanje obaveza, dok je testiranje solventnosti proces procjene adekvatnosti kapitala

za financiranje obaveza grupe poslova. Kriva procjena može dovesti do premalih rezervi što dovodi do insolventnosti, jer sve obaveze ne mogu biti podmirene.

Zato je i kod rezerviranja i kod testiranja solventnosti iznimno bitno dobro predvidjeti odnos obaveza po policama osiguranja i kapitala predviđenog za pokrivanje tih obaveza. U nekim područjima se regresija može koristiti za predviđanje budućih obaveza i za to potrebnih rezervi. Također se može koristiti za pronalazak i usporedbu karakterističnih obilježja stabilnih i nestabilnih tvrtki za testiranje solventnosti.

Poglavlje 2

Osnovna linearna regresija

Osnovna linearna regresija podrazumijeva model koji se sastoji od odzivne varijable Y opisane samo jednom eksplanatornom varijablom X. Razumijevanje tog jednostavnog modela je preduvjet za razumijevanje kompliciranijih modela. Nadalje, osnovna linearna regresija ima prednost da se može lako prikazati na dvodimenzionalnom grafu, što je veoma bitno za stjecanje uvida u podatke.

2.1 Linearni model i njegove pretpostavke

Linearna veza odzivne i jedne eksplanatorne varijable ima sljedeći oblik:

$$Y = \beta_0 + \beta_1 X$$

Iako se možda isprva linearna veza među varijablama čini prejednostavna, u primjeni je veoma korisna te pomoću minimalnih izmjena dobivamo snažan alat za opis veze između odzivne i eksplanatorne varijable.

U praktičnim primjenama često se događa da su dobiveni podaci nasumično raspršeni te stoga ne postoji pravac koji može savršeno opisati vezu među varijablama. Veza između očekivanih vrijednosti varijabli se može dobro opisati, ali je prisutna raspršenost koju pripisujemo slučajnosti te stoga uvodimo sljedeće poboljšanje modela:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \ \forall i$$

gdje su:

- y_i: vrijednost odzivne varijable za i-ti podatak
- x_i: vrijednost eksplanatorne varijable za i-ti podatak

- β_0, β_1: nepoznati parametri, tzv. regresijski parametri koje procjenjujemo na osnovu dostupnih podataka
  - β_0 se naziva odsječak, to je presjek pravca sa y-osi (očekivana vrijednost od Y u slučaju kada je eksplanatorna varijabla $X = 0$)
  - β_1 predstavlja nagib (promjena odzivne varijable Y nastala zbog povećanja vrijednosti eksplanatorne varijable X za jednu jedinicu)
- ϵ_i: greška za i-ti podatak, slučajna razlika između postignute vrijednosti odzivne varijable y_i i njezine očekivane vrijednosti

Kako bismo mogli istražiti matematička svojstva tog modela, moramo napraviti dodatne pretpostavke o distribuciji naših podataka:

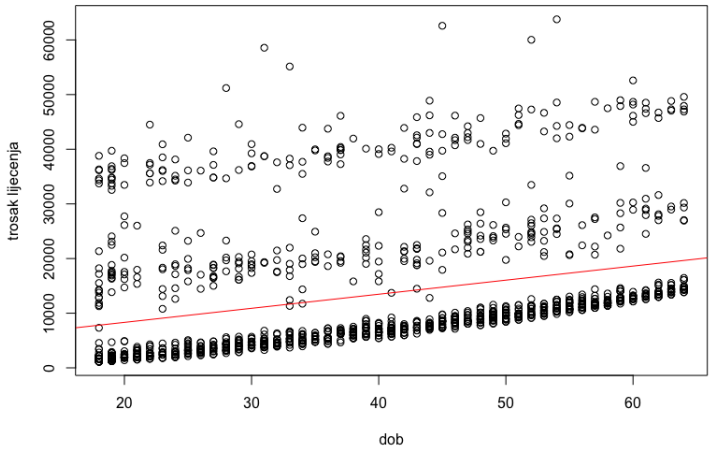
- Prediktori $\{X_1, \ldots, X_n\}$ su nestohastičke varijable.
- $\mathbb{E}[\epsilon] = 0$ i $\text{Var}(\epsilon) = \sigma^2$,
- $\{\epsilon_i\}$ su nezavisne slučajne varijable.
- Greška ϵ ima normalnu razdiobu.

Ove pretpostavke su zasnovane na Gausovskoj teoriji pogreške i takva reprezentacija greške omogućuje konstrukciju mjera kvalitete prilagodbe modela. Valja napomenuti da ove pretpostavke implicitno specificiraju distribuciju naših podataka $\{(X_i, Y_i)\}$ iako je fokus na nepoznatim veličinama $\{\epsilon_i\}$.

Kao primjer osnovne linearne regresije, promotrimo ovisnost medicinskih troškova o dobi klijenta. U ovom slučaju odzivna varijabla y predstavlja trošak liječenja, a eksplanatorna varijabla x predstavlja dob klijenta. Čini se poprilično logično da što je klijent stariji, veća je i vjerojatnost da će se razboliti pa su mu i troškovi veći. Čini se razumnim da je opisani odnos veličina otprilike linearan, a kako troškovi se razlikuju od osobe do osobe jer i ovise o drugim faktorima, modeliramo njihov odnos jednostavnim linearnim modelom sa greškom ϵ:

$$\text{trošak} = \beta_0 + \beta_1 \cdot \text{dob} + \epsilon$$

Na slici 2.1 možemo vidjeti grafički prikaz odnos troškova liječenja u ovisnosti o dobi te prilagođeni regresijski pravac.



Slika 2.1: Prikaz ovisnosti troškova liječenja o dobi za podatke iz poglavlja 7. Crvenom bojom je označen regresijski pravac prilagođen metodom najmanjih kvadrata.

2.2 Procjena koeficijenata metodom najmanjih kvadrata

U slučaju osnovne linearne regresije cilj nam je prilagoditi najbolji mogući pravac danim podacima. Jedan od načina za odrediti traženu regresijsku krivulju je pomoću metode najmanjih kvadrata. U slučaju normalne distribucije greške, metoda najmanjih kvadrata odgovora metodi maksimalne vjerodostojnosti.

Cilj nam je prilagoditi regresijski pravac na način da je suma kvadrata odstupanja između promatranih vrijednosti odaziva y_i i prilagođenih vrijednosti koju predviđa regresijski pravac minimalna za dani skup podataka. U tu svrhu definiramo funkciju koja određuje kvalitetu prilagodbe:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Cilj nam je minimizirati funkciju $Q(.,.)$, a budući da su podaci fiksni, pričamo o minimizaciji s obzirom na regresijske parametre β_0 i β_1. Drugim riječima, trebamo odrediti parametre β_0, β_1 tako da je suma kvadrata reziduala minimalna. Ideja je pronaći parametre za koje su parcijalne derivacije jednake 0:

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = 0 \\ \frac{\partial Q}{\partial \beta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} -2(y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \sum_{i=1}^{n} -2x_i(y_i - (\beta_0 + \beta_1 x_i)) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \sum_{i=1}^{n} x_i(y_i - (\beta_0 + \beta_1 x_i)) = 0 \end{cases}$$

rješavanjem dobivenih jednadžbi dobivamo:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})} \quad \text{i} \quad \hat{\beta_0} = \overline{y} - \hat{\beta_1}\overline{x}$$

gdje je $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, a $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Da se stvarno radi o minimumu vidimo iz sljedećeg:

$$\diamond \quad \frac{\partial Q^2}{\partial^2 \beta_0} = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n} -2(y_i - (\beta_0 + \beta_1 x_i))\right) = 2n > 0$$

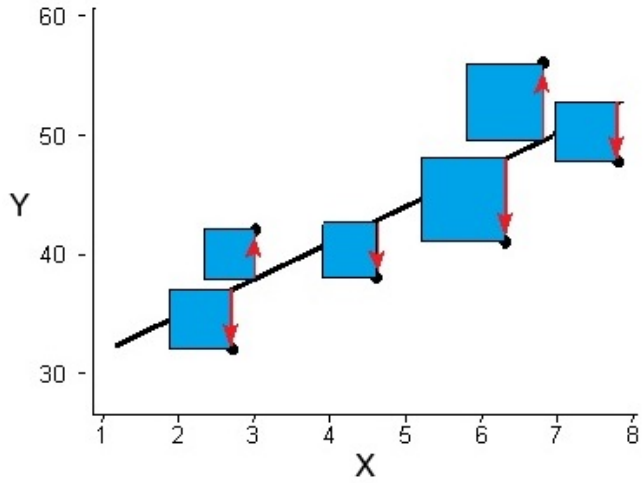
$$\diamond \quad \frac{\partial Q^2}{\partial^2 \beta_1} = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n} -2x_i(y_i - (\beta_0 + \beta_1 x_i))\right) = 2nx_i^2 > 0$$

Sada kad imamo procijenjene parametre modela, možemo pronaći prilagođene vrijednosti koje su dane s:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \ \ \forall i$$

Jedna od primjena ovako dobivenog regresijskog pravca je predviđanje budućih vrijednosti medicinskih troškova.

Možemo se zapitati zašto smo pored raznih mogućnosti minimizacije udaljenosti podataka od regresijskog pravca baš odabrali metodu najmanjih kvadrata. Razlog tome je što je funkcija $Q(.,.)$ koju smo minimizirali diferencijabilna te stoga rješenje postoji, jedinstveno je i može se zapisati u zatvorenoj formi. S druge strane imamo metodu L_1 - regresije koja promatra sumu apsolutnih vrijednosti reziduala no, kao što znamo, apsolutna vrijednost nije neprekidno derivabilna funkcija. Popularnost metode najmanjih kvadrata proizlazi iz činjenice da postoje rezultati o optimalnosti ove metode. Dodatno, pod pretpostavkom normalne distribucije grešaka ϵ_i poznati su točna distribucija procijenjenih koeficijenata i raznih testnih statistika. Grafički prikaz metode najmanjih kvadrata može se vidjeti na slici 2.2.



Slika 2.2: Grafički prikaz prilagodbe regresijskog pravca metodom najmanjih kvadrata. Minimiziramo zbroj kvadrata reziduala, ovdje prikazanih kao plava površina.

Optimalnost metode. Uz pomoć metode najmanjih kvadrata uspjeli smo opisati naše podatke pomoću regresijskog pravca, no bitno je i provjeriti koliko je ta prilagodba zapravo dobra. Naime, gore opisani algoritam se može primijeniti na proizvoljan skup podataka, čak i onaj koji ne karakterizira linearna ovisnost te u takvom slučaju ne dobivamo dobro prilagođeni model. Kako bi gore opisana metoda bila što točnija trebaju biti zadovoljene sljedeće pretpostavke:

1. $\mathbb{E}(\epsilon_i) = 0 \quad \rightarrow$ očekivana greška je 0, tj. nema sistemske greške. To znači da je veza između eksplanatorne i odzivne varijable uistinu linearna.

2. $\text{Var}(\epsilon_i) = \sigma^2 \quad \rightarrow$ raspršenost greški je konstantna

3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j \quad \rightarrow$ nema korelacije među greškama promatranih varijabli, tj. promatrane varijable ne utječu jedna na drugu i ne postoje latentne varijable koje utječu i na x i na y.

4. $\epsilon_i \sim N(0, \sigma^2) \quad \rightarrow$ greške su otprilike normalno distribuirane

Sljedeći teorem nam govori o svojevrsnoj optimalnosti metode najmanjih kvadrata ukoliko su gornje pretpostavke zadovoljene. Dobiveni procjenitelj ima najmanje očekivano kvadratno odstupanje od točne vrijednosti od svih nepristranih linearnih procjenitelja. To dodatno opravdava uporabu metode najmanjih kvadrata.

Gauss - Markovljev teorem:
Ako vrijede svojstva (1) – (3) onda su procjenitelji $\hat{\beta}_0$ i $\hat{\beta}_1$ nepristrani ($\mathbb{E}(\hat{\beta}_0) = \beta_0$ i $\mathbb{E}(\hat{\beta}_1) = \beta_1$). Štoviše, među svim nepristranim linearnim procjeniteljima $\hat{\beta}_0$ i $\hat{\beta}_1$ imaju najmanju varijancu, tj. imaju najmanje očekivano kvadratno odstupanje od točne vrijednosti.

Uočimo da nema zahtjeva o normalnosti distribucije grešaka, no usprkos tome postoje slučajevi u kojima su nelinearni ili pristrani procjenitelji prikladniji od onih dobivenih provođenjem metode najmanjih kvadrata nad podacima s ne-gaussovskim greškama.

Dakle, regresijski koeficijenti su nepristrani u slučaju kada vrijede pretpostavke Gauss-Markovljevog teorema i njihova varijanca je minimalna. Moguće je točno izračunati tu varijancu procjenitelja:

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left\{ \frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right\}$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Vidimo da dobar dizajn eksperimenta može poboljšati kvalitetu procjenitelja.

- Dobit ćemo precizniju regresijsku liniju ako povećamo broj opažanja n.
- Regresijski pravac ima manju varijancu ako su eksplanatorne varijable dobro raspršene.
- Pomoću dobro odabranih procjenitelja varijanca greške σ^2 može biti mala.
- Odječak $\hat{\beta}_0$ je precizniji što je srednja vrijednost uzorka bliže 0.

Ako su greške normalno distribuirane (Gaussovske), tada uz kao gore opisane očekivanje i varijancu imamo poznatu točnu distribuciju procjenitelja. Također, rješenje metode najmanjih kvadrata je procjenitelj najveće vjerodostojnosti Gaussovske greške.

Dodatne prednosti rješenja metode najmanjih kvadrata su:

- Regresijski pravac prolazi kroz središte ravnoteže $(\overline{x}, \overline{y})$.
- Suma reziduala je jednaka 0 tj. srednja vrijednost reziduala je uvijek 0.

2.3 Procjena standardne devijacije grešaka

Osim regresijskih koeficijenata, kako bismo u potpunosti procijenili model, potrebno je procijeniti i varijancu greške koja je nužna za sve testove i intervale pouzdanosti. Procjena je zasnovana na sumi kvadrata reziduala:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Ideja je da ako smo dobro procijenili koeficijente β_0 i β_1, onda će reziduali otprilike odgovarati slučajnim greškama ϵ_i. Kako greške imaju očekivanje nula, a varijancu σ^2, razumno je očekivati da će otprilike slično vrijediti i za reziduale te tada će prosjek kvadrata reziduala otprilike biti jednak σ^2.

Može se pokazati da zbroj kvadrata reziduala $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ima χ^2 distribuciju sa $n-2$ stupnja slobode, pa je stoga $\hat{\sigma}$ nepristran procjenitelj standardne devijacije greške. Intuitivno, n stupnjeva slobode imamo od svih dostupnih podataka, no 2 stupnja izgubimo jer smo prilagođavali parametre β_0 i β_1 takve da su kvadrati reziduala što manji. Sličan fenomen se događa i kad imamo p eksplanatornih varijabli (ako među njih uključimo i odsječak) te je tada procjenitelj varijance grešaka dan sa

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

2.4 Interval pouzdanosti nagiba

Procjena nagiba $\hat{\beta}_1$ je slučajna varijabla, jer je funkcija danih podataka koji su nasumični. Uz zadovoljene pretpostavke metode najmanjih kvadrata (očekivanje je 0, varijance su konstantne, greške su normalno distribuirane i nekorelirane) Gauss-Markovljev teorem nam kaže da će procjenitelj $\hat{\beta}_1$ biti blizu stvarne vrijednosti β_1, ali ne i identičan. Moramo uzeti u obzir da je $\hat{\beta}_1$ izračunat iz uzorka, što znači da bi u slučaju postojanja drugačijeg uzorka, ili kada bismo izostavili neke podatke, dobivena procjena bila drugačija. Sada nam je zadatak tu nesigurnost izraziti u obliku intervala pouzdanosti, koji nam daje više informacija od samog procjenitelja.

Formula je dana s

$$\hat{\beta}_1 \pm qt_{0.975,\,n-2}\,\hat{\sigma}_{\hat{\beta}_1} = \hat{\beta}_1 \pm qt_{0.975,\,n-2}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}},$$

pri čemu je $qt_{0.975,\,n-2}$ predstavlja 95% kvantil Studentove t-distribucije s $n-2$ stupnja slobode. Ova formula se dobije uz pomoć distribucije procjenitelja $\hat{\beta}_1$. Interval pouzdanosti je tako konstruiran da će se u 95% slučajeva pravi nagib β_1 uistinu u njemu i nalaziti.

Kako bismo saznali da li je razumno da vrijednost b bude nagib, provodimo testiranje:

$$H_0: \ \beta_1 = b$$
$$H_1: \ \beta_1 \neq b$$

Testna statistika i njena nul-distribucija dane su s:

$$T_{H_0} = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$

Kako bismo odredili područje prihvaćanja i odbijanja hipoteze, koristimo nul-distribuciju testne statistike, što je Studentova t-distribucija s $n-2$ stupnja slobode, te promotrimo p-vrijednost kako bi odlučili da li prihvaćamo ili odbijamo nul-hipotezu. Lako se pokaže da je regija prihvaćanja jednaka gore navedenom intervalu pouzdanosti te stoga testiranje provodimo tako da samo provjerimo da li se ta vrijednost nalazi u intervalu pouzdanosti.

Interval pouzdanosti odsječka Analogno kao što postoje interval pouzdanosti i test za testiranje značajnosti β_1, tako također postoje i za β_0, no nisu od velikog praktičnog značaja. U praksi je pravilo da se regresija ne provodi bez slobodnog člana, tzv. odsječka. Čak iako nul-hipoteza nije odbijena, te postoji mogućnost da je slobodan član 0, ostavljamo ga obično u modelu, čak i kada teorija nalaže da $x = 0$ povlači $y = 0$.

Izbacivanje odsječka iz modela znači da regresijski pravac prolazi kroz ishodište što je poprilično snažna restrikcija koja često rezultira lošom prilagodbom modela. Najčešći slučaj kad do toga dolazi je kada se vrijednosti prediktora X u trening podacima nalaze daleko od nule te imamo nelinearnu ovisnost između odaziva Y i prediktora X, iako linearni model možda i dobro lokalno opisuje trening podatke.

2.5 Predviđanje

Osnovni cilj linearne regresije je generirati predikciju za odaziv Y na temelju dane vrijednosti eksplanatorne varijable X, tj. ono što očekujemo da će Y biti na temelju danog X. Budući da je $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$, možemo procijeniti vrijednost odaziva u točki x kao

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Jedino je smisleno predviđanje unutar domene eksplanatornih varijabli trening podataka, tj. onih koje smo koristili za prilagodbu modela, te se taj postupak naziva interpolacija. S druge strane, ekstrapolacija predstavlja predviđanje izvan okvira vrijednosti eksplanatorne varijable na trening podacima, što je rizično jer nemamo sigurnost da tamo linearna regresija vrijedi.

Interval pouzdanosti za očekivanje odaziva. Kao što znamo, regresijski koeficijenti su slučajne varijable pa je samim time i regresijski pravac slučajna varijabla, tj. mijenja se u ovisnosti o podacima. Iz tog razloga je bitno razumjeti i znati vizualizirati varijabilnost prilagođenih vrijednosti. Koristeći distribuciju procjenitelja koeficijenata, možemo dobiti formulu za 95% pouzdani interval za očekivanu vrijednost odaziva:

$$\mathbb{E}[Y|X = x] : \hat{\beta}_0 + \hat{\beta}_1 x \pm q t_{0.97.5;\, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

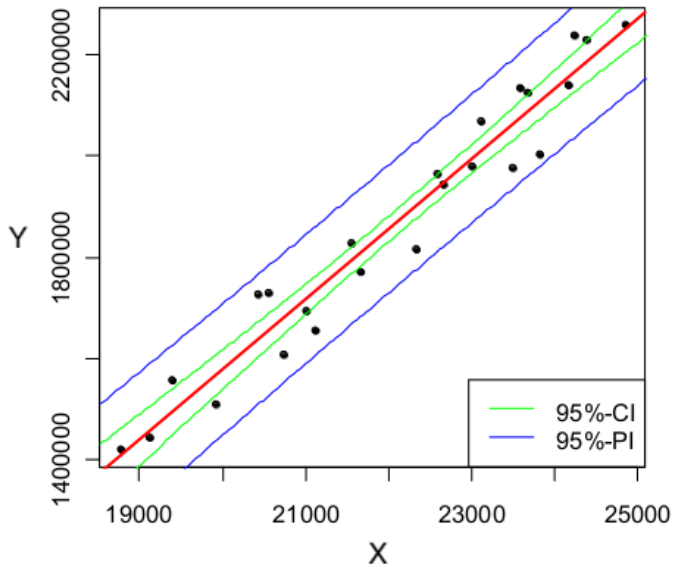
Predikcijski interval. Iako nam gornja formula kazuje nešto o području gdje se nalazi $\mathbb{E}[Y|X = x]$, to nije valjani interval pouzdanosti za sam odaziv $Y|X = x$. Razlog tomu je taj što su promatrane vrijednosti Y dodatno raspršene oko regresijskog pravca zbog greške ϵ, te uzevši to u obzir dolazimo do formule za pouzdani interval za Y:

$$y(x) : \hat{\beta}_0 + \hat{\beta}_1 x \pm q t_{0.97.5;\, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

To možemo zamisliti kao dodatno prošireni interval pouzdanosti za očekivanu vrijednost. Taj interval, koji nam daje područje gdje bi se trebala nalaziti vrijednost odaziva Y, se naziva *predikcijski interval*.

Ilustraciju intervala pouzdanosti prilagođene vrijednosti i predikcijskog intervala možemo vidjeti na slici 2.3. Područje unutar zelenih linija predstavlja interval pouzdanosti regresijskog pravca tj. područje unutar kojeg se otprilike nalazi točan regresijski pravac, dok područje unutar plavih linija predstavlja predikcijski interval

budućih opažanja, tj. gdje otprilike možemo očekivati buduće podatke. Vidimo kako je predikcijski interval širi zbog dodatne nesigurnosti uzrokovane greškom ϵ.

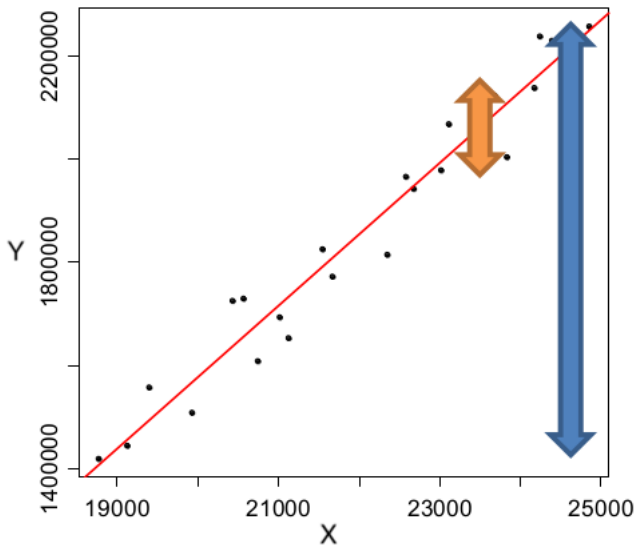


Slika 2.3: Prikaz intervala pouzdanosti za točan regresijski pravac (zeleno) i predikcijskog intervala (plavo) za prilagođeni linearni model (crveno).

2.6 Koeficijent determinacije

U ovom dijelu cilj nam je donijeti neke zaključke o odnosu odzivne i eksplanatorne varijable na temelju indikatora kvalitete i statističkih testova. Intuitivan način za izmjeriti prikladnost modela (eng. goodness-of-fit) je pomoću koeficijenta determinacije koji se označava kao R^2. Koeficijent determinacije mjeri za koliki dio sveukupne varijacije je odgovorna regresijska funkcija, tj. proporcija varijance odzivne varijable koja je objašnjena prediktorskim varijablama.

Kada bismo trebali predvidjeti vrijednost odzivne varijable bez ikakvog saznanja o eksplanatornoj varijabli, najbolji rezultat bismo dobili ako bismo za procjenitelja uzeli prosjek danih mjerenja za odaziv. Međutim, budući da imamo poznata mjerenja za eksplanatornu varijablu, naša predikcija može biti značajno točnija. U slučaju bez eksplanatorne varijable promatramo udaljenost svakog mjerenja odzivne varijable od srednje vrijednosti svih mjerenja, tj. $(y_i - \overline{y})^2$, dok u potonjem slučaju promatramo kvadratno odstupanje od regresijskog pravca: $(y_i - \hat{y}_i)^2$.



Slika 2.4: Ilustracija računanja koeficijenta determinacije R^2 koji mjeri udio varijance odaziva objašnjenog linearnim modelom. Narančasta strelica prikazuje prosječno odstupanje od regresijskog pravca, dok plava strelica prikazuje varijancu odzivne varijable. Koeficijent determinacije se računa uz pomoć omjera tih dvaju veličina.

Ukoliko stvarno postoji nekakav odnos između X i Y, procjena pomoću regresijskog pravca je znatno bolja jer su odstupanja manja. To možemo izmjeriti na sljedeći način:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \in [0, 1]$$

Brojnik predstavlja raspršenost podataka od prilagođenih vrijednosti, tzv. RSS (eng. residual sum of squares). Nazivnik predstavlja raspršenost podataka oko njihove srednje vrijednosti. To je ilustrirano na slici 2.4.

Maksimalna vrijednost koju koeficijent determinacije može postići je 1 i ona predstavlja slučaj kada svi podaci leže na regresijskoj liniji. Drugi ekstrem je 0 i to se događa kada je regresijski pravac dan s $y = \overline{y}$, tj. kada nema eksplanatornih varijabli koje opisuju odzivnu varijablu. Što je R^2 bliže 1, to se veći postotak raspršenosti može objasniti pomoću eksplanatorne varijable.

Teško je odgovoriti na pitanje koja vrijednost koeficijenta determinacije je zadovoljavajuća. Ne postoje općenite smjernice o tome koliko minimalno koeficijent determinacije treba biti kako bi regresija bila uspješna, niti postoje testovi za R^2. Moramo imati na umu da velik R^2 ne znači nužno da je prilagođeni model dobar te su

vrlo često moguća dodatna poboljšanja pomoću alternativnih modela (koristeći neke transformacije). I obrnuto, ako je R^2 mali, ne mora značiti da je prilagođeni model loš. Na primjer, ako je varijanca grešaka σ^2 velika, koeficijent determinacije također ovisi i o varijanci grešaka σ^2; što je ona veća, veći je udio varijance odaziva koji ne možemo objasniti uz pomoć regresije te je stoga i R^2 manji, čak i ako regresijska krivulja savršeno opisuje očekivanje odzivne varijable.

Poglavlje 3

Višestruka linearna regresija

Varijacija vrijednosti odzivne varijable je vrlo rijetko posljedica samo jedne eksplanatorne varijable, čak i u relativno jednostavnim primjerima. Linearni model u kojem je odzivna varijabla opisana pomoću više eksplanatornih dan je s

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon,$$

gdje je odzivna varijabla i dalje označavana sa Y, dok je p različitih eksplanatornih varijabli označeno s $X_1, \ldots, X_p$.

U višestrukoj linearnoj regresiji nemamo samo odnos između jednog prediktora i odzivne varijable, nego možemo imati kompleksniji utjecaj više eksplanatornih varijabli na odzivnu varijablu te čak i međusoban odnos nekoliko prediktora. Također, postaje teško odrediti i koje su eksplanatorne varijable bitne, tj. koje varijable bi stvarno mogle imati utjecaj na odzivnu varijablu te je njihov utjecaj teže vizualizirati.

Osnovna linearna regresija bila nam je poprilično korisna jer nam je omogućavala laku vizualizaciju podataka što je bilo korisno za raspoznavanje odnosa između odzivne i eksplanatorne varijable, određivanje pogodne transformacije varijabli i detektiranje outliera. U slučaju višestruke linearne regresije, prediktori se nalaze u p-dimenzionalnom prostoru, pa stoga ne postoji dvodimenzionalni graf koji može prikazati odnos između tih podataka u punoj općenitosti.

Jedan od načina za pristupiti rješavanju ovog problema je pomoću vizualizacije univarijatnih distribucija svake varijable (odzivne i eksplanatornih) uz pomoć histograma. To je često prvi korak u analizi podataka. Iako nam to ne daje cijelu sliku, omogućuje nam uočavanje asimetričnosti podataka, prisutnost outliera i nekih drugih bitnih karakteristika kao npr. nepotpunost podataka. Kada bismo uočili veliku asimetričnost samo kod nekih varijabli imali bismo dobru naznaku da moramo

izvršiti transformaciju tih varijabli, jer u suprotnom rezultat linearne regresije neće biti reprezentativan.

Drugi tip vizualizacije nam je napraviti graf raspršenosti odzivne i svake eksplanatorne varijable zasebno. Međutim, bitno je napomenuti da taj postupak nije ekvivalentan višestrukoj linearnoj regresiji jer se regresijski koeficijenti kao i njima pripadne p-vrijednosti mijenjaju te se ne može dočarati kompleksni odnos više varijabli. Ta metoda služi samo za vizualizaciju i intuitivno shvaćanje podataka.

Bitno je i napomenuti da će se rezultati regresijske analize moći bolje razumjeti ako su svi podaci u mjernim jedinicama s kojima smo upoznati. Kod linearne transformacije varijable je prednost to što se rezultati regresije ne mijenjaju već samo regresijski koeficijenti i izraz za grešku. Kod nelinearnih transformacija kao npr. logaritam, korijen, kvadrat i slično, regresijska veza i rezultat se mijenjaju, no to nije nužno loše jer to znači da smo prilagodili potpuno novi model.

U nastavku nam je cilj prilagoditi model višestruke linearne regresije podacima, tj. trebamo procijeniti regresijske parametre iz podataka na način da rješenje bude na neki način optimalno. Kao i do sad, tehnika koju ćemo proučavati je minimiziranje sume kvadrata reziduala, tzv. metoda najmanjih kvadrata.

3.1 Metoda najmanjih kvadrata

I u slučaju višestruke linearne regresije nastavljamo koristiti metodu najmanjih kvadrata uz sljedeću funkciju kvalitete prilagodbe:

$$Q(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2$$

Promotrimo općenitu jednadžbu za i-ti podatak u višedimenzionalnom linearnom modelu:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Najelegantniji pristup je zapisati te jednadžbe u matričnom obliku:

$$y = X\beta + E \quad \text{gdje} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{p1} & x_{p2} & \ldots & x_{pp} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{i} \quad E = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

Vektor y nazivamo vektorom odgovora, matricu X matricom dizajna, vektor β vektorom koeficijenata te vektor E vektorom greške. Sada funkciju kvalitete prilagodbe

možemo zapisati kao

$$Q(\beta_0, \ldots, \beta_p) = \|y - X\beta\|_2^2.$$

Za minimum $\hat{\beta}$ funkcije Q vrijedi da je $\nabla Q(\hat{\beta}) = 0$, tj. $\partial_{\beta_i} Q(\hat{\beta}) = 0$. Ako zapišemo to u matričnom obliku, dobivamo:

$$-2X^T(y - X\hat{\beta}) = 0 \quad \Leftrightarrow \quad (X^T X)\hat{\beta} = X^T y$$

Kada je $X^T X$ regularna, rješenje regresijskog problema pomoću metode najmanjih kvadrata je jedinstveno i dano s:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Naravno, jedinstvenost rješenja imamo samo u slučaju da je matrica X punog ranga, tj. kada su eksplanatorne varijable linearno nezavisne. Problem nastaje kada matrica X nije punog ranga. Tada se radi o loše formuliranom modelu te su regresijski koeficijenti djelomično nedefinirani. Kako bismo dobili jedinstveno rješenje nužno je promijeniti izgled modela. Najčešći razlozi koji dovode do singularnosti matrice dizajna su:

- **Duplicirane varijable.** Ako isti podatak koristimo u različitim mjernim jedinicama, npr. udaljenost između gradova u metrima i kilometrima. Te dvije varijable su linearno zavisne, stoga jedna od njih mora biti uklonjena.
- **Cirkularne varijable.** Na primjer, kada bismo imali podatke o prekoračenjima podmirenja obaveza po kreditu i ukupno vrijeme prekoračenja (zbroj prethodno navedenih), imali bismo linearno zavisne prediktore i samim time X ne bi bila punog ranga.
- **Više prediktora nego podataka.** Nužan, no ne i dovoljan uvjet za regularnost matrice $X^T X$, je da je $p < n$, tj. treba nam više podataka nego imamo prediktora, što ima smisla jer je u suprotnom regresija preparametrizirana i nema jedinstveno rješenje.

Distribucija i optimalnost procjenitelja metodom najmanjih kvadrata

Procjenitelj dobiven metodom najmanjih kvadrata možemo zapisati na sljedeći način:

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + E) = \beta + (X^T X)^{-1} X^T E \tag{3.1}$$

Kako greška ima očekivanje $\mathbb{E}[E] = 0$, imamo $\mathbb{E}[\hat{\beta}] = \beta$, što znači da je naš procjenitelj nepristran. Koristeći iste pretpostavke kao u jednostavnom linearnom modelu

1. $\mathbb{E}(\epsilon_i) = 0$
2. $\text{Var}(\epsilon_i) = \sigma^2$
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \ \ \forall \ i \neq j$

možemo iz Gauss-Markovljevog teorema zaključiti da ne postoji nepristran linearni procjenitelj koji bi imao manje kvadratno odstupanje od prave vrijednosti, tj. imamo svojevrsnu optimalnost našeg procjenitelja.

Kao i u slučaju jednostavnog modela, točnost regresijskih koeficijenata ovisi o broju prisutnih podataka. Iako Gauss-Markovljev teorem ne zahtjeva normalnu distribuciju grešaka, kako bismo mogli provesti bilo kakva testiranja koja se temelje na distribucijama procjenitelja, nužno je da imamo nezavisne normalno distribuirane varijable jer tada znamo odrediti distrbuciju procjenitelja. Iz tog razloga uvodimo još jedan dodatan zahtjev:

4. $\epsilon_i \sim N(0, \sigma^2)$ nezavisne i jednako distribuirane

Uz zadovoljena sva četiri uvjeta, iz 3.1 imamo da su procjene regresijskih koeficijenata normalno distribuirane:

$$\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1}) \ \text{ i } \ \hat{y} \sim N(X\beta, \sigma^2 X(X^T X)^{-1} X^T).$$

U slučaju da distribucija greške malo odstupa od normalne, po centralnom graničnom teoremu procjenitelji će asimptotski i dalje imati normalnu distribuciju, stoga se u praksi mala odstupanja mogu tolerirati. Ako je zadatak provesti procjenu koeficijenata, ne moramo previše brinuti o distribuciji greške, no u slučaju zaključivanja na osnovu p-vrijednosti (testiranja), odstupanja se ne bi smjela događati.

Kao što smo već rekli, $\hat{\beta}$ i $\hat{y}$ su nepristrani procjenitelji, a budući da imamo poznatu njihovu distribuciju i matricu kovarijacije, lako možemo odrediti pouzdane intervale i vršiti testiranja. Uz normalnu distribuciju, metoda najmanjih kvadrata nam daje procjenitelje maksimalne vjerodostojnosti (MLE), što nam govori da ne postoje nepristani procjenitelji koji su asimptotski učinkovitiji. Ova tvrdnja je jača od tvrdnje Gauss-Markovljevog teorema, ali zato i zahtijeva dodatno svojstvo Gaussovskih grešaka. Kao i u slučaju jednostavnog modela, tražimo da su pretpostavke za višeparametarsku linearnu regresiju ugrubo zadovoljene, te u slučaju da neka nije zadovoljena, potrebno je izvršiti adekvatnu transformaciju podataka.

Procjena varijance greške. Kako bismo mogli provesti testiranja potrebno je procijeniti varijancu greške σ^2. Prvo procjenjujemo koeficijente te zatim procjenu varijance greške dobijemo skaliranjem sume kvadrata reziduala s prikladnim stupnjevima slobode (broj podataka umanjen za broj procijenjenih parametara), što je dano s:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2.$$

3.2 Test značajnosti procjenitelja

Intervali pouzdanosti regresijskih koeficijenata β_j, $j = 0, 1, \ldots, p$ nam omogućuju način izražavanja nesigurnosti u njihovoj procjeni. Sadrže sve nul-hipoteze $\beta_j = b$ koje pripadni test ne odbije, tj. sve 'razumne' vrijednosti za β_j. Promatrajući distribuciju procjenitelja $\hat{\beta}$, dobivamo sljedeću formulu za interval pouzdanosti:

$$\hat{\beta}_j \pm qt_{0.975;n-(p+1)}\hat{\sigma}_{\hat{\beta}_j} = \hat{\beta}_j \pm qt_{0.975;n-(p+1)}\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}$$

Uz pomoć intervala pouzdanosti možemo testirati je li proizvoljna vrijednost b moguća za regresijski koeficijent β_j:

$$H_0 : \beta_j = b$$
$$H_A : \beta_j \neq b.$$

Ukoliko vrijednost b nije u 95%-intervalu pouzdanosti, tada odbijamo nul-hipotezu na razini 95%. To proizlazi iz činjenice da testna statistika, koja se naziva t-statistika, pod nul-hipotezom slijedi t distribuciju sa $n - p - 1$ stupnjeva slobode:

$$T_{H_0:\beta_j=b} = \frac{\hat{\beta}_j - b}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-p-1}$$

Stoga, ako je b izvan odgovarajućih kvantila te distribucije, odbacujemo nul-hipotezu.

Najvažniji takav test je za slučaj $b = 0$, gdje testiramo da li j-ta eksplanatorna varijabla X_j uopće ima utjecaj na Y ili je $\beta_j = 0$. Ako je odgovarajuća t-statistika prevelika po apsolutnoj vrijednosti, to znači da je vrlo malo vjerojatno da eksplanatorna varijabla u pitanju nije značajna, jer tad bi t-statistika bila blizu 0.

Iako se testiranja poprilično jednostavno provode, interpretacija rezultata nije trivijalna te postoji par situacija koje nas mogu navesti na pogrešne zaključke te stoga valja postupati sa oprezom:

- **Problem višestrukog testiranja.** Ako ponavljamo testiranje za istu razinu značajnosti, ukupna greška prve vrste se povećava. Specifično, za p hipoteza greška iznosi $1-(1-\alpha)^p$. To znači da ako vršimo puno testiranja, nekad ćemo dobiti da je neka varijabla značajna, iako ona to nije.

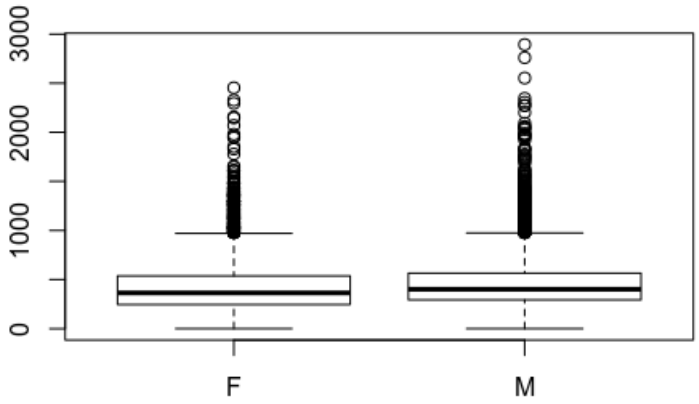
- **Nul-hipoteza nikad nije odbijena.** Iako znamo da u nekim slučajevima eksplanatorne varijable imaju utjecaj na odzivnu, može se dogoditi da test kaže da niti jedna nije značajna. Uzrok tomu je korelacija među eksplanatornim varijablama, što dovodi do preraspodjele predikcije pa se naizgled niti jedna varijabla ne čini bitnom u prisutnosti ostalih. Test značajnosti neke varijable X_i ispituje da li se prilagodba značajno poboljša dodatkom varijable X_i u model. Ukoliko postoje već u modelu varijable korelirane sa X_i, nećemo vidjeti značajno poboljšanje, makar X_i imala utjecaj na Y.

Vrlo je bitno obratiti pozornost i na interpretaciju pojedinog testa: on provjerava utjecaj eksplanatorne varijable X_j na odzivnu varijablu u prisutnosti svih ostalih eksplanatornih varijabli. Bilo kakva (drastična) promjena u eksplanatornim podacima daje drugačiji rezultat. To je važno jer se odluke o izostavljanju varijabli iz modela često donose na temelju rezultata individualnih testova značajnosti. Zbog gore navedenog, ne smije se istovremeno izostaviti više od jedne neznačajne varijable, već se to treba raditi postepeno.

3.3 Kategoričke varijable

Iako za linearni model varijabla odziva uvijek mora biti neprekidna, eksplanatorne varijable mogu, osim neprekidne, biti i kategoričke, gdje postižu neku od konačno mnogo kategorija, kao npr. spol (M/Ž), statusna varijabla (zaposlen/nezaposlen), smjena (jutarnja/popodnevna/noćna), itd. Problem kod kategoričkih varijabli je što one u pravilu nemaju prirodnu mjernu skalu. Iz tog razloga im je potrebno dodijeliti razinu utjecaja na odzivnu varijablu, što radimo pomoću indikatorskih varijabli koje su u kontekstu regresije poznate pod nazivom dummy varijable.

Najjednostavniji model je model u kojem odzivna varijabla Y (neprekidna) ovisi o jednoj kategoričkoj varijabli X, na primjer, iznos police auto osiguranja u ovisnosti o spolu osiguranika. Varijabla Y predstavlja iznos premije, a X je oznaka spola osiguranika. Odnos između opisanih varijabli se uobičajeno prikazuje pomoću boxplota, vidi sliku 3.1. Na grafičkom prikazu vidimo da muškarci u prosjeku plaćaju veće premije nego žene.



Slika 3.1: Usporedba troškova auto osiguranja za žene i muškarce.

Isti odgovor možemo dobiti primjenom regresijske analize Y s obzirom na X. Pritom trebamo biti oprezni, jer regresija koristi samo numeričke podatke, a prediktor X je kategorička varijabla te ga stoga moramo zamijeniti indikatorskom varijablom koja poprima vrijednost 0 ili 1, ovisno o tome radi li se o muškarcu ili ženi:

$$X_1 = \mathbb{1}(\text{klijent je muškarac}) = \begin{cases} 0 & \text{klijent je žena} \\ 1 & \text{klijent je muškarac} \end{cases}$$

Dakle, jednostavan linearan model je dan s:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

gdje β_0 predstavlja očekivanu premiju za žene, a $\beta_0 + \beta_1$ za muškarce. Taj model prilagođava različiti odsječak, ovisno o spolu.

Kada bismo imali više od dvije razine kategoričke varijable, npr. nacionalnost, tada bismo trebali imati $l - 1$ indikatorskih varijabli u modelu, pri čemu je l broj razina kategoričke varijable. Jedna razina je uvijek referentna i odgovara slučaju kad je svih $l - 1$ indikatorskih varijabli jednako nuli. Ne stavljamo svih l indikatorskih varijabli jer bi tada njih l bilo zavisno sa odsječkom (zbroj svih indikatorskih varijabli je jednako 1). Stoga koeficijenti predstavljaju razliku efekta s obzirom na referentnu vrijednost kategoričke varijable.

3.4 Interakcija varijabli

Dodatkom kategoričkih varijabli smo dodavali samo indikatorske varijable u model. Njihov efekt je samo pomicanje prilagođene funkcije za konstantu u ovisnosti o vrijednosti kategoričke varijable. To znači da za različite vrijednosti kategoričke varijable su pravci paralelni. Na primjer, moguće je da su neki koeficijenti linearne regresije različiti za različite razine kategoričke varijable. Tu situaciju je moguće modelirati pomoću samo jedne regresijske jednadžbe koristeći indikatorske varijable.

Na primjer, ako imamo jednu kategoričku varijablu $X_1 = \mathbb{1}$(klijent je muškarac) i jednu numeričku varijablu X_2 koja predstavlja visinu plaće, tada naš model za iznos premije auto osiguranja možemo zapisati kao:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2, + \epsilon$$

gdje smo u model dodali produkt varijabli $X_1 X_2$.

Promotrimo prvo podatke koji se odnose na žene, tj. za koje je $X_1 = 0$:

$$Y = \beta_0 + \beta_2 X_2 + \epsilon$$

Dakle, imamo regresijski pravac s odsječkom β_0 i nagibom β_2. Međutim, model za muškarce dan je sa

$$Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \epsilon,$$

što rezultira regresijskim pravcem s različitim odsječkom $\beta_0 + \beta_1$ i nagibom $\beta_2 + \beta_3$. Tj., vidimo da interakcijski model rezultira dvama regresijskim pravcima s različitim nagibom i odsječkom.

β_1 predstavlja razliku u odsječcima za drukčiji spol u modelu, a β_3 predstavlja posljedičnu promjenu nagiba. Scenarij u stvarnosti u kojemu bi se nagib promijenio je da plaća utječe na iznos premije, jer tada su auti skuplji pa im popravak više košta, no koeficijent promjene je različit za žene i muškarce jer je vjerojatnije da će muškarci voziti prekomjernom brzinom u sportskim autima te tako izazvati sudar i tada je odgovarajući koeficijent nagiba veći za muškarce.

Bitno je još napomenuti da upotreba interakcijskog modela nije ograničena samo na kombinacije neprekidnih i kategoričkih varijabli. Njihova je prednost što se mogu vrlo jednostavno vizualizirati. Interakcijski model je prikladan u slučajevima kada postoji ili sumnjamo da postoji razlika u utjecaju jednog prediktora na odaziv ovisno o vrijednosti koju drugi prediktor (uglavnom kategorički) poprima.

3.5 Predviđanje

Bitna primjena višeparametarske linearne regresije je predviđanje. Kada dobijemo neke vrijednosti eksplanatornih varijabli, čak i ako nisu bile dio formiranja modela, možemo odrediti predviđenu pripadnu odzivnu varijablu:

$$\mathbb{E}[Y_*|x_{*1}, x_{*2}\ , \ \dots\ , x_{*p}] \approx \hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \ldots + \hat{\beta}_p x_{*p}$$

Kod jednostavnog modela rekli smo da je predikcija unutar granica promatranih eksplanatornih varijabli dobra ako regresijski pravac nema sistemsku grešku, no u ovom slučaju je puno teže reći što je unutar a što izvan granica promatranih varijabli jer je poprilično teško vizualizirati p-dimenzionalni prostor. Štoviše, čak i kada se sve nove vrijednosti eksplanatornih varijabli nalaze u blizini prethodno promatranih, nemamo garanciju da nije došlo do ekstrapolacije. Ovaj fenomen je poznat pod nazivom prokletstvo dimenzionalnosti: p-dimenzionalan prostor je velik i čak i kada se nalazimo unutar hiperkocke koja je definirana promatranim eksplanatornim varijablama, nove vrijednosti se mogu nalaziti na području koje podaci za prilagodbu nisu obuhvatili. Međutim, dokle god u modelu nema sistemske greške i dokle god su nove vrijednosti unutar hiperkocke, predviđanja su u pravilu dobra. Osim samih predviđanja, moramo moći znati interpretirati njihovu preciznost, što vidimo iz pouzdanog intervala za očekivanje odaziva i predikcijskog intervala za buduće vrijednosti:

$$\textbf{pouzdani interval: } \hat{y}_* \pm t_{0.975;n-(p+1)}\hat{\sigma}\sqrt{x_*^T(X^TX)^{-1}x_*}$$

$$\textbf{predikcijski interval: } \hat{y}_* \pm t_{0.975;n-(p+1)}\hat{\sigma}\sqrt{1 + x_*^T(X^TX)^{-1}x_*}$$

gdje je $x_*^T = (1, x_{*1}, x_{*2}, \ \dots, x_{*p})$ predikcijski vektor za novo opažanje uključujući slobodni član.

Lako možemo uočiti da je predikcijski interval širi od pouzdanog intervala, jer uz nesigurnost procjene očekivane vrijednosti odgovora $\mathbb{E}[Y_*|x_{*1}, x_{*2}\ , \ \dots\ , x_{*p}]$, imamo i samu nasumičnost odgovora y, čija je stvarna varijanca $\hat{\sigma}$.

Poglavlje 4

Određivanje prikladnosti modela

4.1 Koeficijent determinacije

U slučaju jednostavne regresije, koeficijent determinacije smo interpretirali kao mjeru kvalitete prilagodbe koja uspoređuje raspršenost odzivne varijable sa i bez saznanja o regresijskoj liniji. Iako vizualizacija kao na slici 2.4 zbog većeg broja varijabli ovdje nije moguća, ideja ostaje ista; R^2 izražava koji udio ukupne raspršenosti proizlazi iz regresije te je dan s:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \in [0, 1].$$

Najveća vrijednost je $R^2 = 1$ i postiže se kada se svi podaci nalaze na regresijskoj hiperravnini. Drugi ekstrem, $R^2 = 0$, se postiže kada regresijska prilagodba nimalo ne pomaže u predviđanju odaziva te je tad $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p \approx 0$.

Kao i u slučaju jednostavne regresije, dobiveni koeficijent determinacije trebamo interpretirati s oprezom. Naime, što je veći broj eksplanatornih varijabli, to je manja suma kvadrata reziduala, odnosno R^2 veći. To poboljšanje može biti veće ili manje, ovisno o efikasnosti prediktora, ali iznos statistike R^2 se nikad ne pogorša. To čini koeficijent determinacije, neadekvatnim alatom za uspoređivanje modela s različitim brojem prediktora. To možemo riješiti uvođenjem prilagođenog koeficijenta determinacije:

$$\text{adj}R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

Prilagođeni R^2 je uvijek manji od R^2 jer postoji penalizacija za kompleksne modele (što je veći p to je adjR^2 manji). Razlika je najuočljivija kada ima samo par observacija, puno prediktora i slabi signal. S druge strane, razlika je gotovo zanemariva

kada imamo puno observacija, par prediktora i jaki signal. Generalno se preferira koristiti prilagođeni R^2.

4.2 Testiranje proširenja modela

Testiranje značajnosti prediktora X_j provodimo uspoređujući modele sa i bez X_j i to koristeći t-test. No, postoji i drugi način za usporedbu dva ugniježđena modela. Pretpostavimo da imamo dva modela, od kojih je jedan podskup od drugog:

$$\begin{aligned} \text{Veliki model: } & y = \beta_0 + \beta_1 x_1 + \ \ldots \ + \beta_q x_q + \ \ldots \ + \beta_p x_p \\ \text{Mali model: } & y = \beta_0 + \beta_1 x_1 + \ \ldots \ + \beta_q x_q \end{aligned}$$

Veliki model treba sadržavati iste eksplanatorne varijable koje sadrži i mali jer u suprotnom se ne smatra proširenjem.

Testiramo sljedeće hipotezu:

$$\begin{aligned} H_0 &: \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0 \\ H_A &: \exists\ j \in \{q+1, q+2, \ \ldots \ , p\} : \ \ \beta_j \neq 0 \end{aligned}$$

Usporedba modela se zasniva na usporedbi sume kvadrata reziduala RSS, koja je uvijek manja za veći model jer ima više prediktora. Ako je razlika RSSa između velikog i malog modela premala, nema smisla uzimati prošireni model, no ako je razlika značajna, ima ga smisla razmotriti. Testna statistika je dana s:

$$F = \frac{n-(p+1)}{p-q} \cdot \frac{RSS_{\text{mali}} - RSS_{\text{veliki}}}{RSS_{\text{veliki}}} \sim F_{p-q,\, n-(p+1)}$$

Imamo relativnu usporedbu između prikladnosti modela gdje su u obzir uzeti broj opažanja, ukupan broj eksplanatornih varijabli (prediktora) i razlika u broju prediktora. U slučaju nul-hipoteze, tj. ako dodatni prediktori stvarno nemaju utjecaj na Y, testna statistika ima $F(p-q, n-(p+1))$ distribuciju, te se stoga taj test naziva F-test. Koristeći tu distribuciju možemo odlučiti je li poboljšanje prilagodbe modela značajno veliko ili ne. Valja uočiti i da ako se testiranje proširenja modela svodi na usporedbu modela koji se razlikuju u jednoj eksplanatornoj varijabli, onda se svodi na pojedinačno testiranje značajnosti varijable koje je različito od t-testa opisanog prije.

4.3 Testiranje značajnosti modela - globalni F test

Globalni F test je specijalan slučaj testiranja proširenja modela te je dan na sljedeći način:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_A : \exists j \in \{1, 2, \ \ldots \ , p\} : \ \beta_j \neq 0$$

Tj. nul hipoteza tvrdi da je model potpuno beznačajan. Ovdje se zapravo radi o testiranju proširenja modela gdje uspoređujemo dani model s najjednostavnijim modelom, gdje imamo samo odsječak: $Y = \beta_0 + \epsilon$. Uzimamo uobičajenu testnu statistiku $F(p - q, n - (p + 1))$, no uočimo da je ovdje $q = 0$, tj., u manjem modelu nemamo eksplanatornih varijabli:

$$F = \frac{n - (p + 1)}{p} \cdot \frac{RSS_{\text{mali}} - RSS_{\text{veliki}}}{RSS_{\text{veliki}}} \sim F_{p,\, n-(p+1)}$$

4.4 Dijagnostički dijagrami

U ovom ćemo dijelu proučiti kako provjeriti da li pretpostavke višestrukog linearnog modela vrijede. Razlog za to je što želimo biti sigurni da su procjene vjerodostojne. Drugi razlog je taj što nam razumijevanje grešaka može pomoći u unapređenju modela. Također, kako bi se mogli uzdati u p-vrijednosti i ostale tvrdnje o distribuciji raznih statistika, uvjeti za višestruki linearni model moraju biti zadovoljeni.

Prisjetimo se zahtjeva za provođenje metode najmanjih kvadrata:

- $\mathbb{E}[\epsilon] = 0$
- $\text{Var}(\epsilon) = \sigma^2$
- $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \ \ \forall i \neq j$
- $\epsilon_i \sim \text{N}(0, \sigma^2 I)$ nezavisne

Prve tri pretpostavke su nužne kako bi se mogla provesti procjena najmanjih kvadrata te za validaciju prilagođenih vrijednosti, dok je četvrti uvjet potreban kako bi se mogla provoditi testiranja, pronaći intervali pouzdanosti te predikcijski intervali.

Reziduali kao procjenitelji za greške

Da bi provjerili pretpostavke linearnog modela, trebali bi znati iznos grešaka za svaki podatak. Greške ϵ_i su slučajne varijable jednake razlici $y_i - X_i\beta$ između promatranih varijabli i očekivanih vrijednosti. Problem je što nam stvarni koeficijenti β nisu poznati pa su nam i samim time stvarne greške nepoznate, no tome možemo doskočiti tako što ćemo ih procijeniti rezidualima:

$$r_i = y_i - X_i\hat{\beta}$$

Međutim, moramo biti oprezni jer $\epsilon_i \neq r_i$, štoviše, greška i rezidual imaju i različita svojstva. Čak i kada je varijanca greške konstantna, reziduali će biti blago korelirani i heteroskedastični (varijanca im nije konstantna) zbog grešaka pri prilagodbi modela.

Zapišimo reziduale na sljedeći način:

$$R = y - \hat{y} = y - X\hat{\beta} = y - X(X^TX)^{-1}X^Ty := y - Hy = (I - H)y$$

Iz toga možemo dobiti distribuciju reziduala:

- $\mathbb{E}[R] = \mathbb{E}[y - X\hat{\beta}] = 0$
- $\text{Var}[R] = \text{Var}((I - H)y) = (I - H)\text{Var}(y)(I - H)^T = (I - H)\sigma^2$
  $\Rightarrow \text{Var}(R_i) = (1 - H_{ii})\sigma^2$
- $R_i \sim N(0, (1 - H_{ii})\sigma^2)$ kao linearna kombinacija normalno distribuiranih odzivnih varijabli

Problem nastaje zbog činjenice da su reziduali blago heteroskedastični te ne zadovoljavaju u potpunosti svojstva koja bi greške trebale zadovoljavati. Što je podatak i udaljeniji od ostalih podataka, to je i H_{ii} veći pa je i varijanca reziduala manja. Stoga, kako bismo umanjili svojstva heteroskedastičnosti, često se promatraju **standardizirani reziduali**, gdje se svaki rezidual skalira sa njegovom procijenjenom standardnom devijacijom:

$$\tilde{r}_i = \frac{r_i}{\hat{\sigma}\sqrt{1 - H_{ii}}},$$

gdje je $\hat{\sigma}$ procjena standardne devijacije grešaka.

Ako bi $\hat{\sigma}$ dolazio iz prilagodbe linearnog modela na preostalih $n - 1$ podataka, tada govorimo o studentiziranim rezidualima. Oni su teži za računanje, no precizniji su u prisutnosti outliera te slijede $t_{n-(p+1)}$ distribuciju.

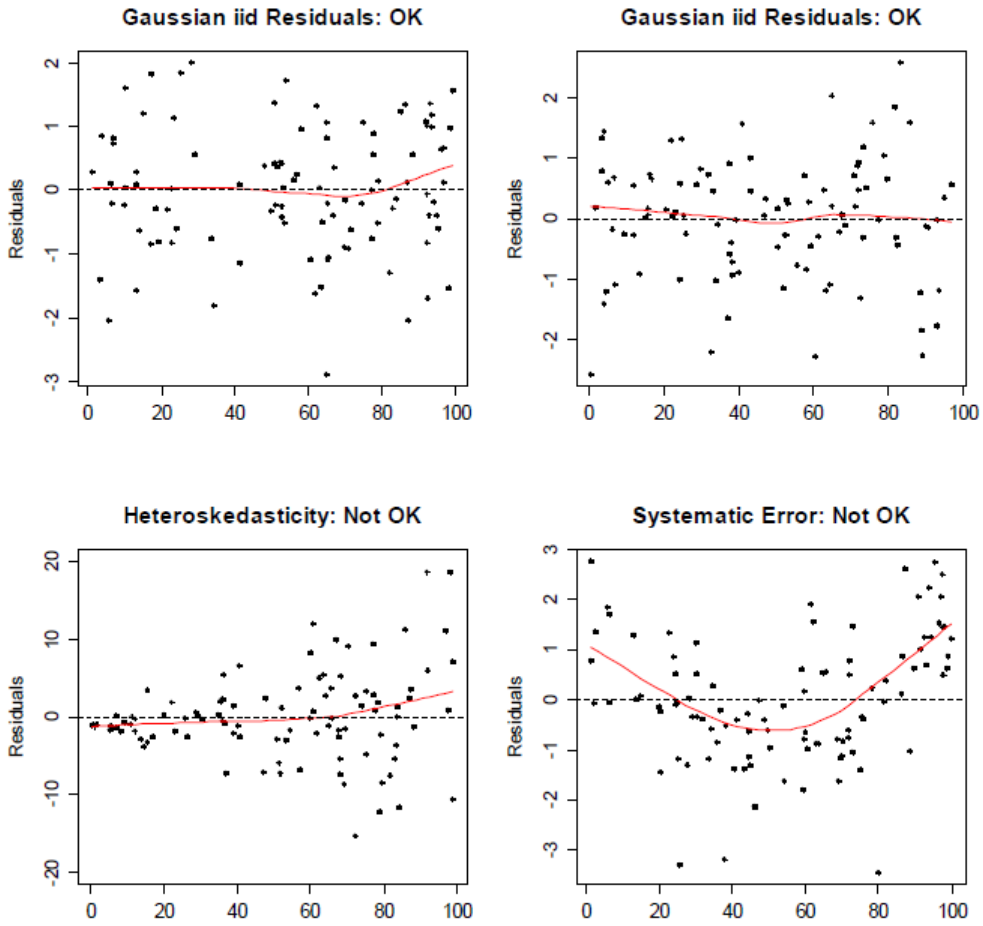
Tukey - Anscombe graf

Tukey - Anscombe graf je najbitniji dijagnostički alat za utvrđivanje modela višestruke linearne regresije. Taj graf prikazuje odnos reziduala r_i i prilagođenih vrijednosti $\hat{y}_i$ te uglavnom služi za ispitivanje da li je očekivana pogreška 0, tj. $\mathbb{E}[\epsilon_i] = 0$, na temelju čega se donosi odluka o ispravnosti modela i stvaranju nepristranih procjenitelja. Moramo biti svjesni da nije moguće provjeriti da je očekivana greška 0 za svaku pojedinu grešku jer će neki reziduali biti veliki a neki jako mali, no nama to ne smeta. Nama je samo bitno da možemo utvrditi je li lokalni prosječni rezidual znatno različitiji od 0, što se nikad ne bi smjelo dogoditi. Na graf se stoga dodaje i regresijska krivulja kako bi lakše odokativno procijenili lokalni prosjek reziduala. Graf možemo lako nacrtati uz pomoć funkcija iz programa R, no kompleksniji dio zadatka je znati pravilno intrepretirati dobiveni graf.

Kako bi pretpostavka o nultoj grešci vrijedila, zahtijevamo da se regresijska krivulja ne udaljava sistematično od x-osi, tj. da je udaljavanje od x-osi slučajno. Ako se regresijska krivulja znatno udaljava od x-osi, regresijski model ima sistemsku grešku te je pretpostavka linearnosti našega modela vrlo vjerojatno narušena. U tom slučaju ne generiramo predikcije, niti provodimo testiranja, već moramo na neki način poboljšati model uz pomoć raznih transformacija. Možemo pokušati napraviti logaritamsku transformaciju podataka, što će nam u većini slučajeva popraviti rezultat.

Za ilustraciju promotrimo sliku 4.1. Na gornje dvije slike vidimo primjer gdje se regresijska krivulja ne udaljava sistematično od x-osi, tj. reziduali su nezavisni i normalno distribuirani te se svaka devijacija pripisuje slučajnosti. Na donje dvije slike je situacija drugačija. Na donjoj desnoj slici imamo sistematično odstupanje od nule te stoga linearnost modela ne vrijedi. S druge strane, na donjoj lijevoj slici se očekivana greška ne udaljava sistematično od 0, no što je prilagođena vrijednost veća, veća je i varijanca reziduala. To narušava uvjet homoskedastičnosti reziduala koji je potreban kako bi se metoda najmanjih kvadrata mogla primijeniti.

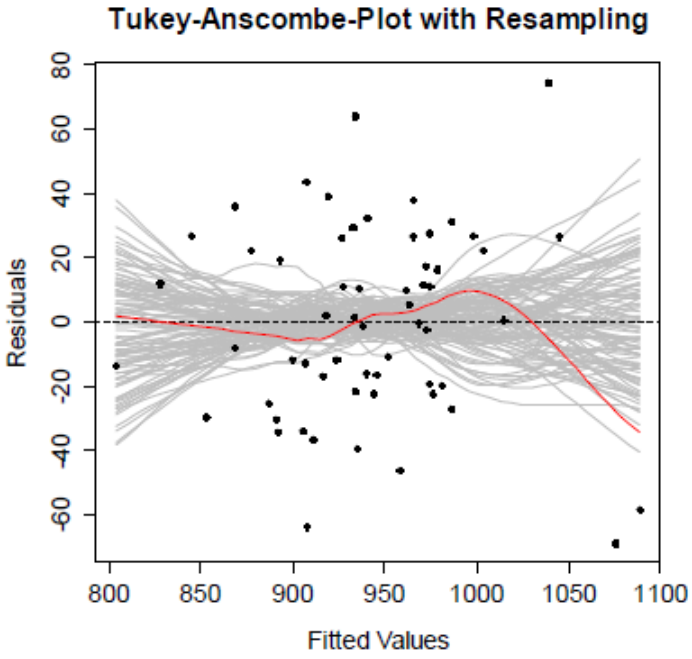
Radi li se o sistemskom udaljavanju regresijske krivulje od x-osi ili o slučajnosti je odluka koju donosi stručnjak, no mi ovom problemu možemo doskočiti tehnikom ponovljenog uzorkovanja. Metoda se svodi na zadržavanje starih prilagođenih vrijednosti $\hat{y}_i$ kojima pridružimo nove vrijednosti reziduala. Nove vrijednosti reziduala su generirane iz skupa starih reziduala uz mogućnost ponavljanja. Odgovarajuće regresijske krivulje za novodobivene podatke $(\hat{y}_i, r_i^*)$ dodamo na Tukey-Anscombe graf u obliku sivih linija i taj postupak ponovimo stotinjak puta. U ovom postupku se neće pojaviti sistematično odstupanje regresijske krivulje od nule jer smo reziduale na slučajan način pridruživali prilagođenim vrijednostima. Zapravo nam te nove



Slika 4.1: Različite mogućnosti izgleda Tukey-Anscombe dijagnostičkog grafa. Gornje dvije slike ne ukazuju na probleme, dok donje dvije slike ukazuju na problem heteroskedastičnosti (lijevo) i sistematične greške (desno).

krivulje predočavaju kolika devijacija od x-osi može biti rezultat slučajnosti.

Na slici 4.2 vidimo situaciju u kojoj je originalna regresijska krivulja (predstavljena crvenom linijom) poprilično dobro raspoređena unutar područja u kojem je krivulja rezultat slučajnosti (predstavljeno sivim linijama), izuzev na desnom rubu gdje je na samoj granici, no budući da su u pitanju samo dvije vrijednosti s jako negativnim rezidualima, možemo taj problematični dio zanemariti i zaključiti da je odstupanje slučajno.



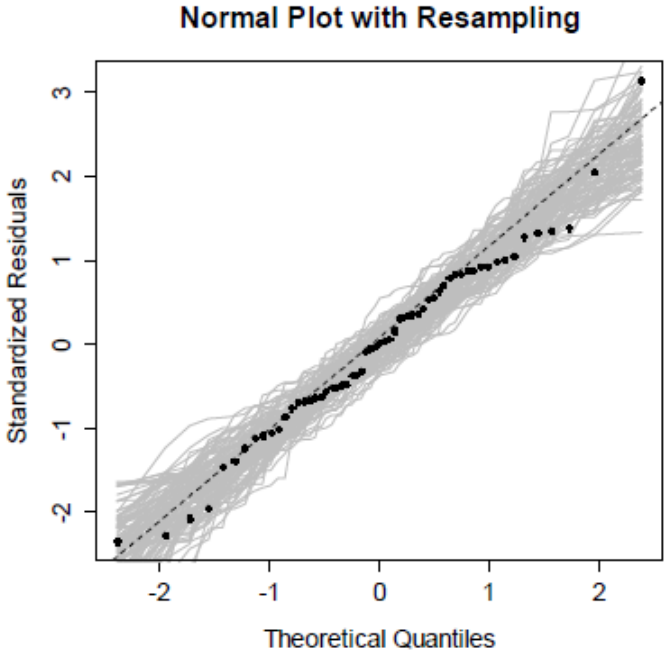
Slika 4.2: Da li regresijska krivulja za dane podatke (crveno) značajno odstupa od x-osi, možemo zaključiti uspoređujući sa odgovarajućim krivuljama za slučajno generirane reziduale (sivo).

Q-Q graf

Pomoću Q-Q grafa možemo ispitati jesu li greške zaista normalno distribuirane slučajne varijable tako što nacrtamo sortirane standardizirane reziduale u odnosu na kvantile standardne normalne razdiobe. Budući da se normalna distribucija greške prenosi na reziduale, graf ne bi trebao pokazivati znakove sistematične devijacije u odnosu na pravac provučen kroz prvi i treći kvantil danih distribucija. Ukoliko to nije slučaj, tada vjerujemo da ne vrijedi da su reziduali nezavisni i jednako normalno distribuirani. Također valja napomenuti da se odstupanje na grafu može dogoditi i u slučaju heteroskedastičnih podataka, bez obzira da li su oni normalno distribuirani.

Kada se nacrta graf, diskutabilno je odrediti kada je devijacija slučajna, a kada je sistematična. Kao i u prethodnom poglavlju, odluku o tome nam može olakšati tehnika ponovljenog uzorkovanja. Nacrtamo 100 slučajnih uzoraka duline n iz normalne distribucije koji imaju isto očekivanje i standardnu devijaciju kao reziduali, pa ako nam svi uzorci upadnu u područje određeno sa 100 slučajnih

uzoraka, zaključujemo da nema sistematičnog odstupanja od normalne distribucije. Primjer možemo vidjeti na slici 4.3.

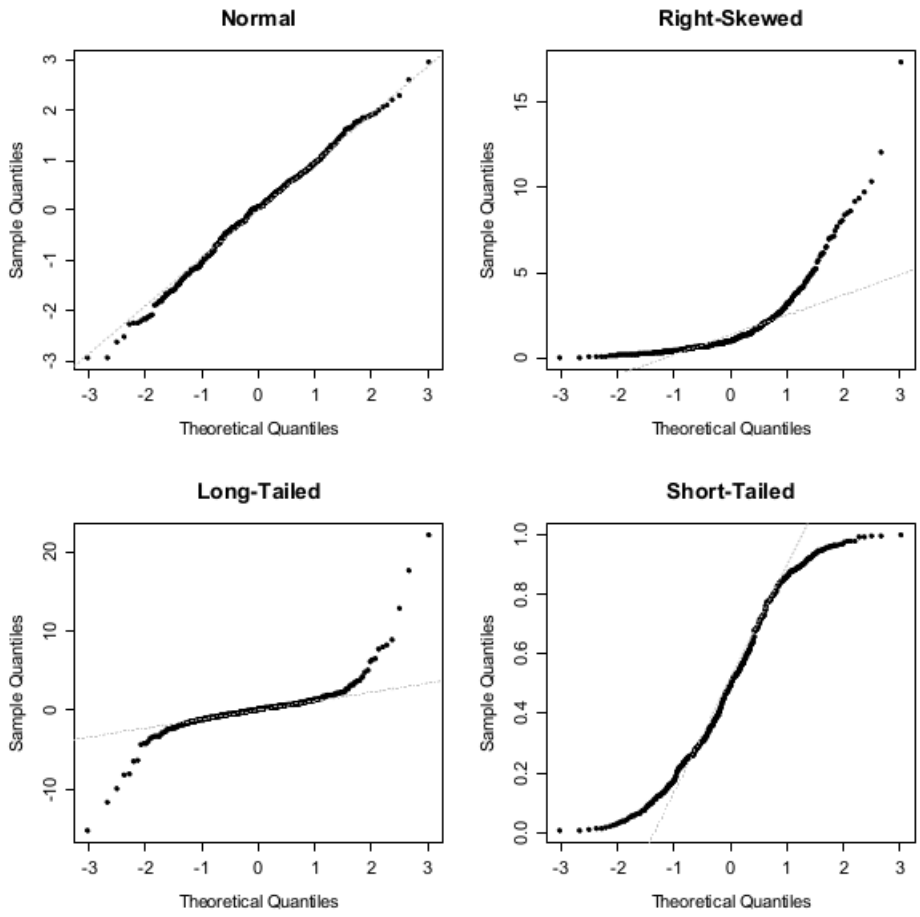


Slika 4.3: Da li nacrtani podaci značajno odstupaju od pravca $y = x$ možemo zaključiti uspoređujući sa odgovarajućim krivuljama za slučajno generirane reziduale (na slici sivo).

U slučaju da devijacija od pravca $y = x$ postoji, ovisno o vrsti odstupanja reziduala poduzimamo različite mjere. Procjena pomoću metode najmanjih kvadrata je iznimno osjetljiva na nakošene reziduale (eng. skewed) iz razloga što se oni često podudaraju sa sistematičnom greškom modela koja i narušava Tukey-Anscombe graf. Distribucije reziduala koje su laganih ili teških repova ne predstavljaju problem za određivanje prilagođenih vrijednosti i koeficijenata dokle god su simetrične, jer gotovo uopće nemaju utjecaj na prilagođene vrijednosti koje su i dalje vjerodostojne i nepristrane. Međutim, narušena je točnost određivanja intervala pouzdanosti, jer što je rep distribucije teži, to je točnost manja jer distribucija više odstupa od normalne. Za teške repove bi trebalo proširiti intervale pouzdanosti jer velika vjerojatnost pojavljivanja ekstremnih vrijednosti povećava raspršenost raznih statistika.

Ilustraciju raznih slučajeva možemo vidjeti na slici 4.4. Prvi graf prikazuje tipičnu normalnu distribuciju te ne uočavamo odstupanje od pravca, dok je drugi graf najproblematičniji jer predstavlja situaciju s nakošenom distribucijom. Treći i

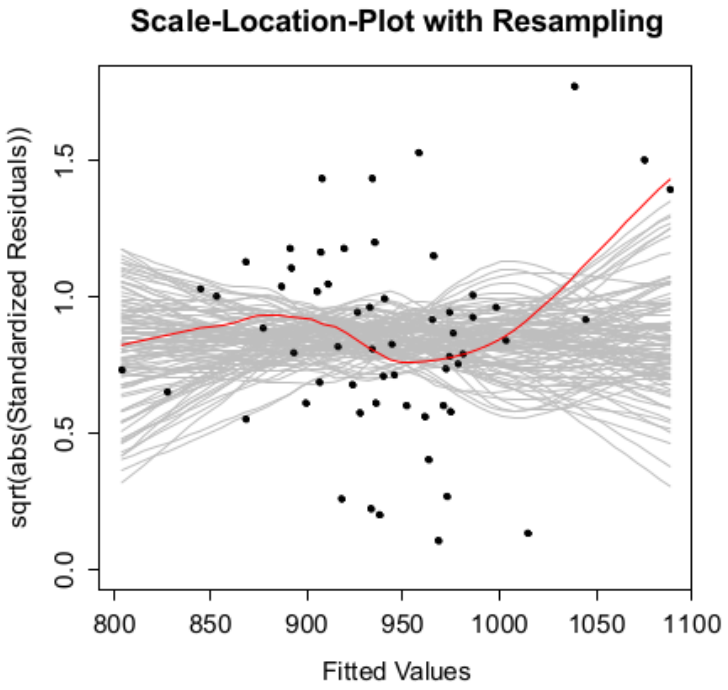
četvrti graf prikazuju situacije lakog i teškog repa reziduala, koje nisu baš idealne, ali simetričnost distribucije reziduala djelomično umanjuje problem jer, iako prilagođene vrijednosti nisu idealne, vjerojatnije je da su nepristrane nego u slučaju nakošene distribucije reziduala.



Slika 4.4: Moguć izgled QQ-dijagnostičkih dijagrama. Slika gore lijevo ne pokazuje nikakvo odstupanje distribucije reziduala od normalne, dok preostale tri pokazuju razna odstupanja: nakošenost (gore desno), teški repovi (dolje lijevo), laki repovi (dolje desno).

Konstantnost varijance grešaka

Graf skale i položaja (eng. scale-location plot) nam služi za detektiranje nekonstantne varijance, tj. heteroskedastičnosti. Već smo spomenuli da se heteroskedastičnost može iščitati iz Tukey-Anscombe grafa, ali ovaj graf je učinkovitiji jer na njemu vidimo korijen apsolutne vrijednosti standardiziranih reziduala u odnosu na prilagođene vrijednosti. Krucijalna stvar je uzeti apsolutnu vrijednost. Na graf dodajemo regresijsku krivulju, koja će u slučaju nepostojanja heteroskedastičnosti biti otprilike horizontalna. Također, možemo detektirati sistematično odstupanje tako što iz naših podataka generiramo slučajan uzorak $(\hat{y}_i, r_i^*)$ te nacrtamo odgovarajuće regresijske krivulje radi usporedbe.



Slika 4.5: Primjer izgleda skala-položaj dijagnostičkog grafa, koji služi za detekciju heteroskedastičnosti. Da li je regresijska krivulja za dane podatke (crveno) dovoljno paralelna sa x-osi, možemo zaključiti uspoređujući sa odgovarajućim krivuljama za slučajno generirane reziduale (sivo).

Ilustraciju možemo vidjeti na grafu 4.5. Na grafu izgleda kao da varijanca raste kako su prilagođene vrijednosti veće, no tome možemo pokušati doskočiti primjenom logaritamske transformacije podataka.

Poglavlje 5

Regresijski modeli za određivanje premija osiguranja

U ovom dijelu proučit ćemo regresijske primjene izračuna premija kod određivanja faktora povjerenja i u bonus-malus sustavima iskustvenog određivanja premija osiguranja. Sustavi iskustvenog određivanja premija (eng. experience rating system) obuhvaćaju službene metode za uključivanje novih podataka o štetama u obnovljene premije kratkoročnih ugovora (npr. automobilsko osiguranje ili osiguranje od ozljede).

5.1 Klasifikacija rizika i iskustveno određivanje rizika

Klasifikacija rizika je ključna za određivanje premija osiguranja. Osiguravatelji prodaju pokrića po cijenama koje su dovoljne za podmiriti troškove očekivanih potraživanja, administrativne troškove i očekivani profit koji bi trebao kompenzirati trošak kapitala potrebnog za podršku prodaje pokrića. U modernom svijetu tržište osiguranja je razvijeno i konkurentno, što potiče osiguravatelje da klasificiraju rizike kojima se izlažu kako bi dobili adekvatne premije. Ta se klasifikacija temelji na poznatim obilježjima osiguranika.

Kao primjer promotrimo tvrtku koja prodaje osiguranje automobila u slučaju kvara. Izračunavamo premiju za dva klijenta koji imaju ista obilježja (mjesto stanovanja, dob, spol, godišnji prihod, itd.) izuzev zanimanja: jedan klijent je vozač taksija te se služi osobnim automobilom kao izvorom prihoda, a drugi klijent koristi automobil isključivo za privatne potrebe. Iz iskustva znamo da će osoba koja koristi automobil isključivo u privatne svrhe imati manja potraživanja od osobe koja ga

koristi u poslovne svrhe i u skladu s time se donosi odluka od iznosu premije, jer u suprotnom će neka druga osiguravajuća kuća ponuditi bolju premiju i izgubit ćemo sigurnijeg klijenta. To natjecanje među osiguravajućim kućama vodi određivanju premija na temelju opažajnih karakteristika što je poznato pod nazivom klasifikacija rizika. U kontekstu regresijskog modeliranja gledamo na to kao na modeliranje distribucije potraživanja u ovisnosti o raznim eksplanatornim varijablama.

Međutim, često osiguravajuće kuće ne koriste samo općenite podatke o klijentima, nego prilagođavaju visinu premije ponaosob. U većini slučajeva se premija određuje i na temelju veze između osiguravatelja i osiguranika koja se s vremenom razvije. To omogućuje osiguravatelju da odluku o cijeni premije donese na temelju neprimijećenih karakteristika osiguranika uzevši u obzir i prethodna iskustva šteta. Modeliranje premija uz poznatu povijest šteta se naziva **iskustveno određivanje premije**.

Iskustveno određivanje premija se provodi ili retrospektivno ili prospektivno. Kod retrospektivnih metoda osiguraniku se u slučaju okolnosti povoljnih za osiguravatelja daje povrat dijela premije. Retrospektivne premije su česte kod životnih osiguranja gdje osiguranici često dobijaju bonus po isteku ugovora. Kod osiguranja imovine i osiguranja od nesreće se češće koriste prospektivne metode, gdje se "dobre" osiguranike nagrađuje nižim novim premijama.

Dvije prospektivne metode prikladne regresijskim modelima su metode za određivanje faktora povjerenja i bonus-malus metoda. Bonus-malus metode se koriste uglavnom u Europi i Aziji, ali su ograničene na osiguranja vezana uz automobile, dok metode procjene povjerenja imaju puno širi spektar upotrebe. Ideja je koristiti iskustvo šteta kako bi se poboljšala klasifikacija osiguranika.

5.2 Metode procjene povjerenja

Metode procjene povjerenja obuhvaćaju metode za određivanje premija osiguranja s primjenom u zdravstvu, nekretninama, nezgodama i životnom osiguranju [3]. Ideja je na osnovu poznate povijesti šteta i dodatnih informacija razviti formulu za određivanje cijene premije, npr.

$$\text{nova cijena premije} = \zeta \cdot \text{iskustvo štete} + (1 - \zeta) \cdot \text{stara cijena premije},$$

gdje koeficijent $\zeta \in [0, 1]$ nazivamo faktor povjerenja.

Slučaj $\zeta = 1$ odgovara potpunom povjerenju i jedino se iskustvo štete koristi za određivanje nove premije. Slučaj $\zeta = 0$ se smatra nepovjerenjem te se u tom slučaju

iskustvo štete zanemaruje i nova premija se određuje isključivo na temelju vanjskih informacija o osiguraniku.

Potpuni i parcijalni faktor povjerenja

Zanima nas kada s pouzdanjem možemo koristiti metodu procjene povjerenja. Kada se ova metoda razvijala, osnovni cilj je bio razlikovati slučaj velikog poslodavca s velikom količinom podataka, dovoljnom za učenje iz iskustva, i malih poslodavaca s nedovoljnom količinom podataka za učenje iz iskustva. Statistički rečeno, cilj je napraviti predviđanje očekivanog potraživanja na temelju iskustva i dobivenih podataka.

Najjednostavniji slučaj je da pretpostavimo da imamo nezavisne, jednako distribuirane troškove $y_1, y_2, \dots y_n$ s očekivanjem μ i standardnom devijacijom σ. Kako bismo dobili potpuni faktor povjerenja, tražimo dovoljan broj podataka n kako bi vrijedilo:

$$\mathbb{P}\left((1-r)\mu \leq \overline{y} \leq (1+r)\mu\right) \geq p$$

za neko malo dopušteno odstupanje r, te uz normalne procjene dobivamo

$$n \geq \left(\frac{\Phi^{-1}\left(\frac{p+1}{2}\right)\sigma}{r\mu}\right)^2.$$

Najmanja vrijednost n za koju je gornja nejednakost zadovoljena predstavlja minimalan potreban broj podataka za pouzdanu procjenu faktora povjerenja.

Vrlo često u praksi aktuari nemaju dovoljno podataka, pogotovo kada su u pitanju manji klijenti, ali čak i u slučaju velikih klijenata može se dogoditi da je od interesa promatrati samo manje podskupove podataka (podijeljeno po spolu, dobi i slično). Za takve grupe podataka se uobičajeno koristi težinsko vagani prosjek iskustva troškova. Uz pretpostavku približne normalnosti podataka, djelomični faktor povjerenja se određuje na sljedeći način:

$$\text{nova premija} = Z \cdot \overline{y} + (1 - Z) \cdot \text{manualna premija}$$

gdje $Z \in [0, 1]$ nazivamo koeficijent povjerenja definiran kao:

$$Z = \min\left\{1, \sqrt{\frac{n}{n_F}}\right\},$$

gdje je n broj podataka a n_F minimalan broj podataka potreban za ostvarivanje potpunog faktora povjerenja.

Što imamo više podataka, to je Z bliže 1, tj. veće grupe podataka su vjerodostojnije te veća je i uloga iskustva $\overline{y}$ u modelu. Analogno, što je koeficijent povjerenja manji (bliže 0), veća je uloga ručne procjene očekivane štete u modelu, tj. karakteristika grupe.

Faktor povjerenja najveće točnosti

Promatramo troškove $y_1, \ldots, y_n$ male grupe te želimo procijeniti srednju vrijednost za tu grupu. Budući da je grupa mala, prosjek $\overline{y}$ ne predstavlja nužno najbolju procjenu. Pretpostavimo i da imamo vanjsku procjenu prosjeka troškova M, o kojoj razmišljamo kao o ručnoj procjeni štete, dobivenu nekom regresijskom metodom. Zanima nas možemo li kombinirati $\overline{y}$ i M kako bismo pronašli što boljeg procjenitelja.

Ako pretpostavimo da postoje tzv. strukturne varijable koje označavamo s α. To su neopaženi faktori, koji su zajedničke svim opažanjima iz grupe, i uvjetno na α pretpostavljamo da je skup $\{y_1, \ldots, y_n\}$ skup nezavisnih slučajnih jednakodistribuiranih varijabli. Iako su strukturne varijable neopažene, nešto o njima možemo zaključiti iz ponavljajućih opažanja troškova. Za svaku grupu uvjetno očekivanje i varijancu označujemo s $\mathbb{E}(y|\alpha)$ i $\mathrm{Var}(y|\alpha)$, te mi želimo odrediti najbolju procjenu za $\mathbb{E}(y|\alpha)$. Ta procjena se naziva Bühlmanova premija povjerenja [4] i dana je s:

$$\text{nova premija} = \zeta \cdot \overline{y} + (1 - \zeta) \cdot M$$

gdje je ζ koeficijent vjerodostojnosti definiran kao

$$\zeta = \frac{n}{n + \text{omjer}} \quad \text{gdje je} \quad \text{omjer} = \frac{\mathbb{E}(\mathrm{Var}(y|\alpha))}{\mathrm{Var}(\mathbb{E}(y|\alpha))}$$

Upravo ta procjena premije ima najmanju varijancu u klasi svih linearnih i nepristranih prediktora. Ovaj rezultat je zasnovan na pretpostavci nezavisnosti i jednake distribucije troškova.

5.3 Bonus-Malus

Bonus-Malus metode iskustvenog određivanja rizika imaju široku primjenu kod auto osiguranja [5]. Promotrimo prvo određivanje premije zasnovano na promatranim obilježjima (spol i dob vozača, vrsta automobila, mjesto stanovanja, itd.). Primjena samo spomenutih obilježja rezultira a priori premijom, što predstavlja osnovu premije u Americi i Kanadi. Iskustveno određivanje rizika u priču ulazi preko naknada za nesreće i zabilježenih prometnih prekršaja.

Bonus-Malus metoda nam daje bolju integraciju iskustvenih troškova u model. Uobičajeno se osiguranici svrstavaju u nekoliko kategorija. Čim uđu u sustav, svrsta ih se u određenu kategoriju. Nakon godinu dana, ako osiguranik nije bio u nesreći, dobiva "bonus" te je premješten u višu kategoriju. Sukladno tome, ako je došlo do nesreće, osiguraniku se dodjeljuje "malus" te ga se spušta u niže kategorije. Kategorije određuje tzv. Bonus-Malus faktor koji pomnožen s a priori premijom daje a posteriori premiju.

5.4 Regresijsko modeliranje potraživanja

Čest način modeliranja potraživanja je i uz pomoć regresije [6, 2]. Tada pokušavamo na temelju podataka o štetama odrediti vezu između šteta i raznih eksplanatornih varijabli. Određivanje faktora povjerenja i očekivanih potraživanja u okviru regresijskih modela ima više prednosti:

- Mogućnost prilagodbe velikog broja različitih modela, jer regresija ima široku primjenu.
- Dodatna metoda za objašnjavanje procesa određivanja premija.
- Analiza podataka postaje znatno jednostavnija, jer imamo matematičke alate kojima možemo opisati podatke. Također možemo matematički kvantificirati varijabilnost i rizičnost.
- Olakšana grafička analiza i vizualizacija podataka.

Kao što smo i mogli vidjeti na primjeru iskustvenog određivanja premije, regresijska metoda određivanja premije može služiti kao samo jedna komponenta metodologije za računanje premija. Uz pomoć regresije možemo izračunati kolika bi trebala biti premija za prosječnog klijenta koji posjeduje dane karakteristike te je stoga ona stoga i odlična polazna točka i za individualno prilagođene premije. U nastavku ćemo vidjeti dvije takve regresijske analize na primjeru zdravstvenog i auto osiguranja.

Poglavlje 6

Primjer regresijske analize: auto osiguranje

U ovom poglavlju ćemo pokazati primjer regresijske analize na podacima o automobilskom osiguranju za Sjedinjene Američke Države. U skupu podataka je za svakog klijenta zabilježeno potraživanje od auto osiguranja. Želimo pronaći model koji će dobro opisati o čemu ovise troškovi. Na raspolaganju imamo podatke o $n = 9134$ klijenata, pri čemu je za svakog klijenta zabilježeno sljedeće:

- mjesto stanovanja - kratica savezne države
- vrsta osiguranja - tip osiguranja (Basic, Extended, Premium)
- obrazovanje - razina obrazovanja osiguranika (High school or below, College, Bachelor, Master, Doctor)
- status zaposlenja - zaposlen, nezaposlen, bolovanje, umirovljenik ili osoba s invaliditetom
- spol - M/Ž
- prihod - godišnji prihod osiguranika u USD
- bračni status - u braku, slobodan, razveden
- iznos mjesečne premije za automobil - iznos u USD
- vrijeme proteklo od posljednje štete izraženo u mjesecima
- broj polica osiguranja
- razlog potraživanja odštete - sudar, ogrebotina ili udubljenje, tuča i drugo
- ukupna šteta izražena u USD
- vrsta automobila osiguranika (auto s 4 vrata, auto s 2 vrata, luksuzan auto, luksuzan SUV, sportski auto, SUV)
- veličina automobila (veliki, srednji, mali)
- namjena vozila - u osobne ili poslovne svrhe

Prvo ćemo promotriti model u kojem predviđamo troškove na temelju vrste osiguranja, obrazovanja, spola, bračnog statusa, uzroka štete, mjesečne premije za automobil, vremena od posljednje štete i vrste vozila. U R-u to implementiramo na sljedeći način:

```
fit <- lm(Total.Claim.Amount ~ Coverage + Education
        + Gender + Marital.Status + Monthly.Premium.Auto +
        Months.Since.Last.Claim + Claim.Reason + Vehicle.Class,
        data = df)
```

U R-u postoji jedna vrlo korisna funkcija pod nazivom "summary()". Ona kao argument može primiti objekt koji vraća funkcija "lm", koja prilagođava linearni model podacima, te ispisuje mnogo informacija o prilagođenom linearnom modelu koje su nam potrebne za analizu danih podataka [7]. Kada primijenimo spomenutu funkciju na našem linearnom modelu, dobivamo sljedeće informacije:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1168.98  -112.85    25.86    97.10  1714.62

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -94.5095    23.2595  -4.063 4.88e-05 ***
CoverageExtended              -12.0955     8.2926  -1.459 0.144713
CoveragePremium                -4.5031    17.5670  -0.256 0.797694
EducationCollege               -4.7722     5.6990  -0.837 0.402405
EducationDoctor               -55.5518    12.0652  -4.604 4.19e-06 ***
EducationHigh School or Below  43.8017     5.7517   7.615 2.89e-14 ***
EducationMaster               -68.3368     8.7004  -7.854 4.47e-15 ***
GenderM                        35.4930     4.4032   8.061 8.54e-16 ***
Marital.StatusMarried         -22.6118     6.4224  -3.521 0.000432 ***
Marital.StatusSingle          135.7033     7.1024  19.107  < 2e-16 ***
Monthly.Premium.Auto            5.3052     0.3238  16.383  < 2e-16 ***
Months.Since.Last.Claim         0.1228     0.2180   0.563 0.573338
Claim.ReasonHail               -6.6607     5.2786  -1.262 0.207042
Claim.ReasonOther             -20.4684     7.5263  -2.720 0.006549 **
Claim.ReasonScratch/Dent       -9.5284     6.5556  -1.453 0.146125
Vehicle.ClassLuxury Car        44.5698    46.1217   0.966 0.333893
Vehicle.ClassLuxury SUV        11.0614    45.6982   0.242 0.808745
Vehicle.ClassSports Car       -33.2061    17.5033  -1.897 0.057841 .
Vehicle.ClassSUV               -8.2820    15.2104  -0.544 0.586113
```

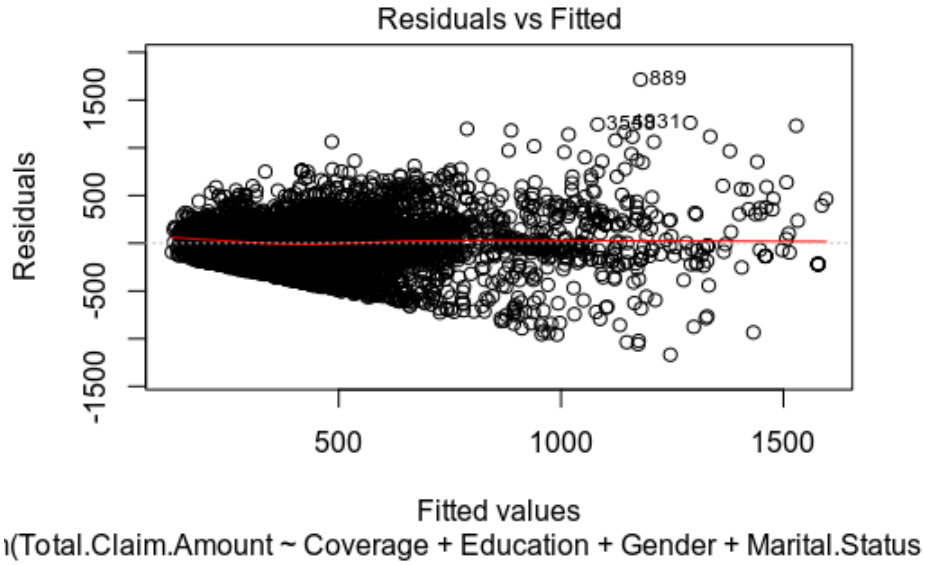
```
Vehicle.ClassTwo-Door Car       0.2396     5.7342   0.042 0.966670
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.5 on 9114 degrees of freedom
Multiple R-squared:  0.4809,Adjusted R-squared:  0.4798
F-statistic: 444.4 on 19 and 9114 DF,  p-value: < 2.2e-16
```

Prvo što možemo promotriti su reziduali koji su rastavljeni u 5 točaka: minimum prvi kvartil, medijan, treći kvartil i maksimum. Minimum i maksimum nam ovdje nisu od pretjeranog interesa, no ostala tri podatka mogu biti prilično poučna. Naime, linearna regresija pretpostavlja normalnu distribuciju reziduala s očekivanjem 0. Nije nužno problem ako distribucija grešaka nije točno normalna, ali bi barem trebala biti simetrična: srednja vrijednost bi trebala biti blizu 0 i apsolutne vrijednosti prvog i trećeg kvartila bi trebale biti približne. Iako nam se može činiti da ne zadovoljavamo niti jedan od tih uvjeta, u stvarnosti prilagođeni model nije u potpunosti loš, ako uočimo da baratamo s vrlo velikim podacima: najveća vrijednost iznosi 1714.62 a najmanja -1168.98, pa promatrano na tako velikoj skali su naznake o prihvatljivosti ovog modela ipak zadovoljavajuće.

Na slici 6.1 možemo vidjeti Tukey-Anscombe graf reziduala. Vidimo da je regresijska krivulja veoma blizu nule, što pokazuje kako reziduali imaju očekivanje veoma blizu nule, što i treba vrijediti za linearni model. Ali s druge strane, kao što znamo, jedna od pretpostavki modela linearne regresije je i homoskedastičnost reziduala. No na grafu raspršenosti uočavamo da su reziduali heteroskedastični, tj. kod predviđanja većih vrijednosti potraživanja štete se događaju i veća odstupanja, što nije toliko nerazumno.

Normalnost grešaka ispitujemo uz pomoć grafa na kojem uspoređujemo standardizirane reziduale s teoretskim kvantilima normalne razdiobe, tzv. QQ graf, što je prikazano na slici 6.2. Lako možemo uočiti da podaci dolaze iz distribucije težeg repa od normalne te pretpostavka o normalnosti ne izgleda zadovoljena. Međutim, to je čest slučaj u praksi jer je vjerojatnost velikih odstupanja puno veća nego što bi bio slučaj kod normalnih grešaka. To može biti zbog nepoznatih faktora koje nemamo uključene u podacima, a bitni su za dani problem.

Kada promotrimo koeficijente modela uočavamo da vrsta osiguranja nije od pretjeranog značaja za ovaj model, kao niti vrijeme proteklo od zadnje štete. S druge strane, čini se da je podatak o obrazovanju klijenta značajan samo kod visokoobrazovanih ljudi pa ćemo tu varijablu zadržati u modelu. Iz istog razloga u modelu ostavljamo uzrok štete i vrstu vozila. Kako je analiza grafa normalnosti ukazala



Slika 6.1: Tukey-Anscombe graf za prilagođeni linearni model na podacima o auto osiguranju.

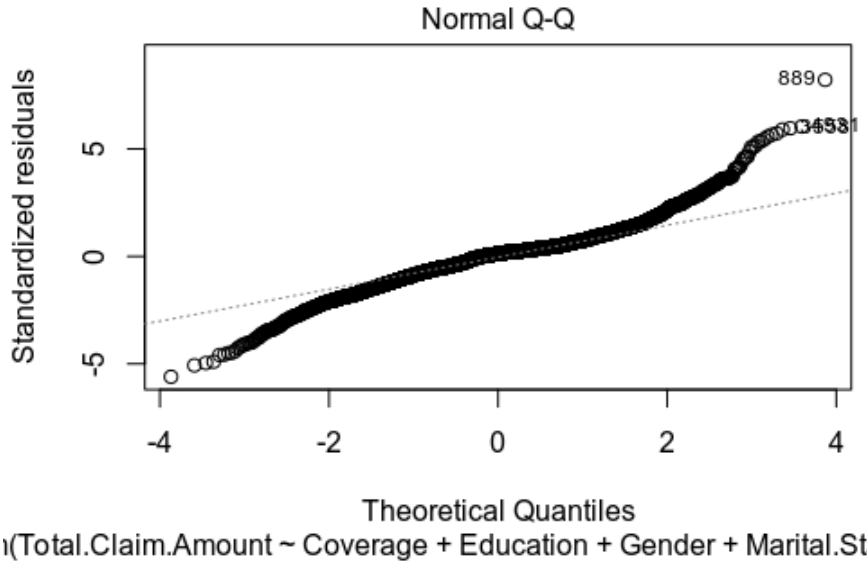
da nam podaci dolaze iz distribucije težeg repa nego normalna, ova analiza nije najtočnija te se treba uzimati s oprezom. Međutim, ona nam daje dobre smjernice i relativna usporedba značajnosti prediktora još uvijek ima smisla.

Nakon micanja varijabli koje su utvrđene kao beznačajne, novi model smo implementirali na sljedeći način:

```
fit2 <- lm(Total.Claim.Amount ~ Education + Gender + Marital.Status +
           Monthly.Premium.Auto + Claim.Reason,  data = df)
```

te najbitnije informacije o modelu dobivamo pozivom funkcije summary:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1128.26  -111.16    26.78    97.04  1720.45

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -97.22788    9.34650 -10.403  < 2e-16 ***
```



Slika 6.2: QQ graf za prilagođeni linearni model na podacima o auto osiguranju.

```
EducationCollege               -5.45418    5.69772  -0.957 0.338463
EducationDoctor               -57.45174   12.06552  -4.762 1.95e-06 ***
EducationHigh School or Below  42.84545    5.75177   7.449 1.03e-13 ***
EducationMaster               -69.59256    8.70381  -7.996 1.45e-15 ***
GenderM                        35.84314    4.40185   8.143 4.37e-16 ***
Marital.StatusMarried         -22.67649    6.41819  -3.533 0.000413 ***
Marital.StatusSingle          136.17888    7.10246  19.173  < 2e-16 ***
Monthly.Premium.Auto            5.28341    0.06412  82.401  < 2e-16 ***
Claim.ReasonHail               -5.88602    5.26983  -1.117 0.264054
Claim.ReasonOther             -18.52463    7.50547  -2.468 0.013600 *
Claim.ReasonScratch/Dent       -8.48467    6.55488  -1.294 0.195558
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.7 on 9122 degrees of freedom
Multiple R-squared:  0.4793,Adjusted R-squared:  0.4787
F-statistic: 763.4 on 11 and 9122 DF,  p-value: < 2.2e-16
```

Analiza reziduala je ponovno veoma slična, no sada imamo drugačiju situaciju s varijablama modela. Uzrok troška i obrazovanje su i dalje dosta upitni. Iz tog

razloga prvo ćemo iz modela maknuti uzrok troška pa usporediti dobiveni model s ovim koristeći χ^2-test.

```
fit3 = lm(Total.Claim.Amount ~ Education + Gender + Marital.Status +
        Monthly.Premium.Auto, data = df)
```

Dva modela uspoređujemo pozivom funkcije anova:

```
anova(fit3, fit2, test='Chisq')
```

i rezultat je sljedeći:

```
Analysis of Variance Table

Model 1: Total.Claim.Amount ~ Education + Gender + Marital.Status
+ Monthly.Premium.Auto

Model 2: Total.Claim.Amount ~ Education + Gender + Marital.Status
+ Monthly.Premium.Auto + Claim.Reason

  Res.Df       RSS Df Sum of Sq Pr(>Chi)
1   9125 401591102
2   9122 401301866  3    289236  0.08677
```

Dakle, varijabla uzrok troška nije značajna.

Još smo jedino sumnjali na obrazovanje, stoga ponavljamo isti postupak, ali sada provjeravamo je li model koji sadrži informaciju o obrazovanju pojedinca (uz prethodno uklonjenu informaciju o vrsti vozila) značajan za određivanje troška:

```
fit4 = lm(Total.Claim.Amount ~ Gender + Marital.Status +
          Monthly.Premium.Auto, data=df)
```

Usporedbom modela vidimo da je od značaja zadržati informaciju o obrazovanju pojedinca na razini značajnosti od 5%:

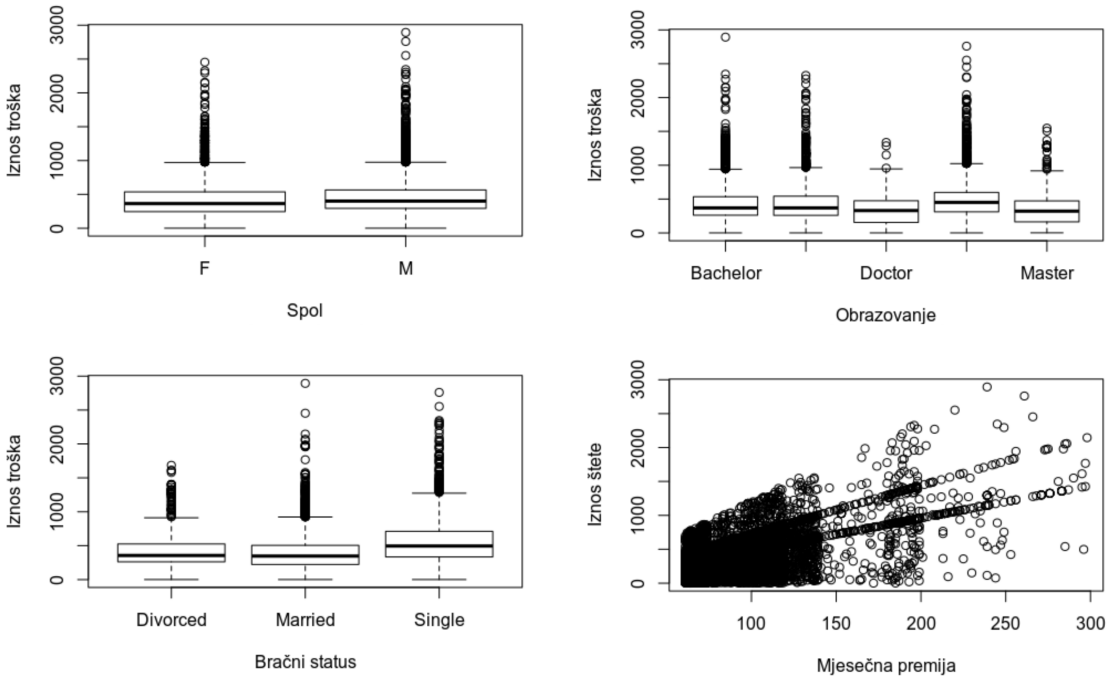
```
Model 1: Total.Claim.Amount ~ Gender + Marital.Status
+ Monthly.Premium.Auto
Model 2: Total.Claim.Amount ~ Education + Gender + Marital.Status
+ Monthly.Premium.Auto
  Res.Df       RSS Df Sum of Sq  Pr(>Chi)
```

```
1   9129 410989831
2   9125 401591102  4   9398729 < 2.2e-16 ***
```

Dakle, od svega što smo vidjeli, najbolji je model u kojem trošak predviđamo na osnovu spola, obrazovanja, bračnog statusa i mjesečne premije za automobil. Iz priložene analize vidimo da linearan model nije idealan, budući da reziduali nisu normalno distribuirani, ali s druge strane taj model može biti koristan, pogotovo kao polazišna točka za neke kompleksnije modele. Koeficijent determinacije R^2 nam kaže da model uspijeva objasniti 47.9% rasipanja, što nije ni loše.

```
    Call:
lm(formula = Total.Claim.Amount ~ Education + Gender + Marital.Status +
    Monthly.Premium.Auto, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1124.32  -110.86    26.85    96.99  1718.57

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -101.78127    9.00928 -11.297  < 2e-16 ***
EducationCollege               -5.87741    5.69632  -1.032 0.302198
EducationDoctor               -58.74074   12.05146  -4.874 1.11e-06 ***
EducationHigh School or Below  41.84085    5.73736   7.293 3.29e-13 ***
EducationMaster               -70.51358    8.69799  -8.107 5.86e-16 ***
GenderM                        35.23701    4.39584   8.016 1.23e-15 ***
Marital.StatusMarried         -24.44566    6.36691  -3.839 0.000124 ***
Marital.StatusSingle          136.10576    7.10368  19.160  < 2e-16 ***
Monthly.Premium.Auto            5.29563    0.06383  82.961  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.8 on 9125 degrees of freedom
Multiple R-squared:  0.479,Adjusted R-squared:  0.4785
F-statistic:  1048 on 8 and 9125 DF,  p-value: < 2.2e-16
```

Na slici 6.3 možemo otprilike i uočiti da postoji neka veza između potraživanja štete te spola, obrazovanja i bračnog stanja te se stoga rezultati naše regresijske analize čine smisleni. Vidimo da muškarci u prosjeku čine više štete, no ona se u prosjeku smanjuje što je čovjek više obrazovan. Na slici 6.3 možemo vidjeti i ovisnost štete o iznosu premije. Vidimo da su te dvije varijable pozitivno korelirane



Slika 6.3: Ovisnost medicinskih troškova o spolu (gore desno), obrazovanju (gore lijevo), bračnom stanju (dolje lijevo) te dosadašnjoj mjesečnoj premiji (dolje desno) za podatke o štetama za auto osiguranje.

te stoga vidimo kako za određivanje buduće premije valja uzeti u obzir dosadašnju premiju.

Poglavlje 7

Primjer regresijske analize: zdravstveno osiguranje

U nastavku ćemo proučiti primjer regresijske analize, gdje ćemo analizirati ovisnost medicinskih troškova o raznim faktorima. Na raspolaganju imamo podatke o $n = 1338$ osoba. Za svaku osobu znamo sljedeće informacije:

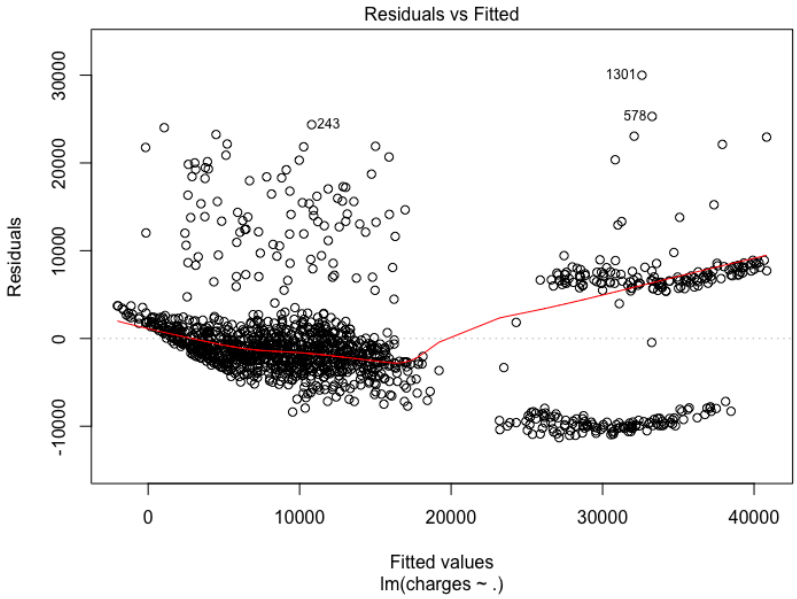
- godišnje medicinske troškove izražene u USD
- spol
- dob
- indeks tjelesne mase (BMI - body mass index) = težina osobe izražena u kilogramima podijeljena sa kvadratom visine u metrima
- podatak o broju djece
- je li osiguranik pušač ili ne
- regija stanovanja - jedna od 4 kategorije: northeast, northwest, southeast, southwest

Koristeći dane podatke, želimo predvidjeti prosječan iznos medicinskih troškova kako bismo odredili kolika će biti premija koju bi osiguranik trebao plaćati.

Prvo prilagodimo linearni model ovisnosti iznosa premije o svim faktorima:

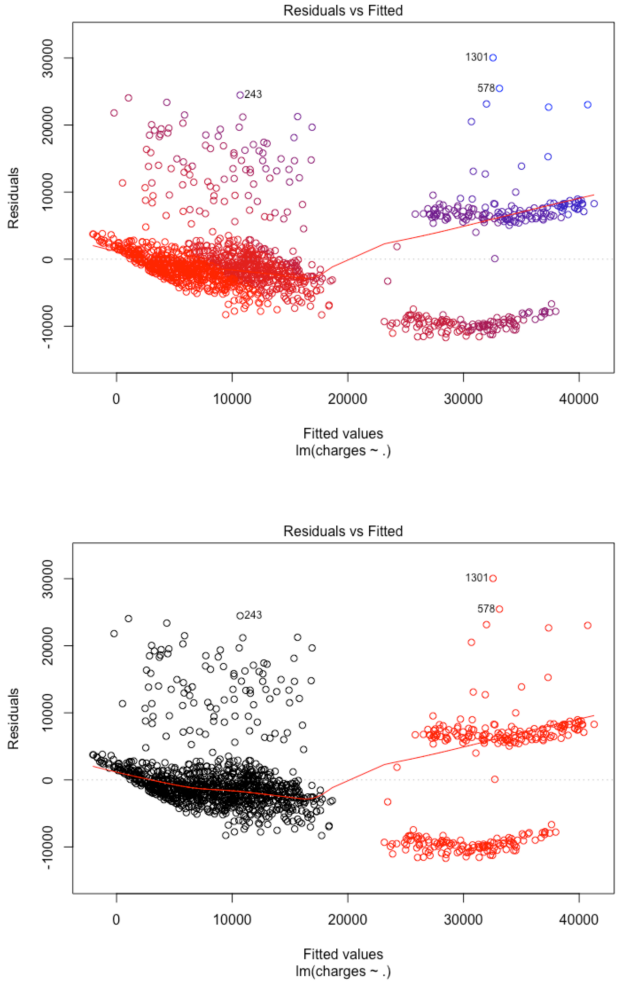
```
fit <- lm(charges~., data=df)
summary(fit)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
gendermale        -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```



Slika 7.1: Tukey-Anscombe graf za prilagođeni linearni model gdje koristimo sve podatke.

Na slici 7.1 vidimo Tukey-Anscombe graf za naš prilagođeni linearni model. Vidimo da otprilike reziduali imaju očekivanje nula. Međutim, možemo uočiti kako varijanca reziduala raste za velike prilagođene vrijednosti. Također, vidimo da se podaci nalaze u nekoliko grupacija, što ukazuje na postojanost dodatne strukture distribucije. Ako proučimo na dijagnostičkim dijagramima gdje se nalaze točke u ovisnosti o parametrima, možemo primijetiti uzorak.



Slika 7.2: Tukey-Anscombe graf prikazan na slici 7.1, gdje su točke obojane u ovisnosti o indeksu tjelesne mase (slika gore, pri čemu plavo znači veliki indeks tjelesne mase) i u ovisnosti o tome da li je osoba pušač (slika dolje, pri čemu crvena boja označava pušače).

Koristeći kod

```
plot(fit, 1, col=df$smoker)
plot(fit, 1, col=colorRampPalette(c('red','blue'))(10)
                            [cut(as.numeric(df$charges), breaks=10)])
```

dobivamo dijagrame na slici 7.2. Možemo uočiti da se pušači nalaze desno na slici, tj. za njih je predviđeni trošak velik, kao što smo mogli i očekivati, a da grupacija gore desno na slici odgovara osobama sa velikim indeksom tjelesne mase. Stoga ćemo pokušati dodati interakcijski član između varijabli 'pušač' i 'indeks tjelesne mase':

```
fit2 <- lm(charges~ . + smoker:bmi, data=df)
summary(fit2)

Call:
lm(formula = charges ~ . + smoker:bmi, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-14580.7  -1857.2  -1360.8   -475.7  30552.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2223.454    865.611  -2.569  0.01032 *
age                263.620      9.516  27.703  < 2e-16 ***
sexmale           -500.146    266.518  -1.877  0.06079 .
bmi                 23.533     25.601   0.919  0.35814
children           516.403    110.179   4.687 3.06e-06 ***
smokeryes       -20415.611   1648.277 -12.386  < 2e-16 ***
regionnorthwest   -585.478    380.859  -1.537  0.12447
regionsoutheast  -1210.131    382.750  -3.162  0.00160 **
regionsouthwest  -1231.108    382.218  -3.221  0.00131 **
bmi:smokeryes     1443.096     52.647  27.411  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4846 on 1328 degrees of freedom
Multiple R-squared:  0.8409,Adjusted R-squared:  0.8398
F-statistic:   780 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Vidimo da je interakcijski član veoma značajan, što znači da pretilost (visoki indeks tjelesne mase) ima različit utjecaj na medicinske troškove, ovisno o tome je li osoba pušač ili nepušač. Također vidimo da je pozitivan za kategoriju pušača, što znači da se pušačima troškovi povećavaju još i jače u ovisnosti o pretilosti nego nepušačima.

Promotrivši koeficijente vidimo da podaci o spolu i mjestu stanovanja pokazuju naznake neznačajnosti. Kako bismo to provjerili napravit ćemo prvo model u kojem uklanjamo varijablu o mjestu stanovanja:

```
fit3 <- lm(charges~ . + smoker:bmi - region, data=df)
```

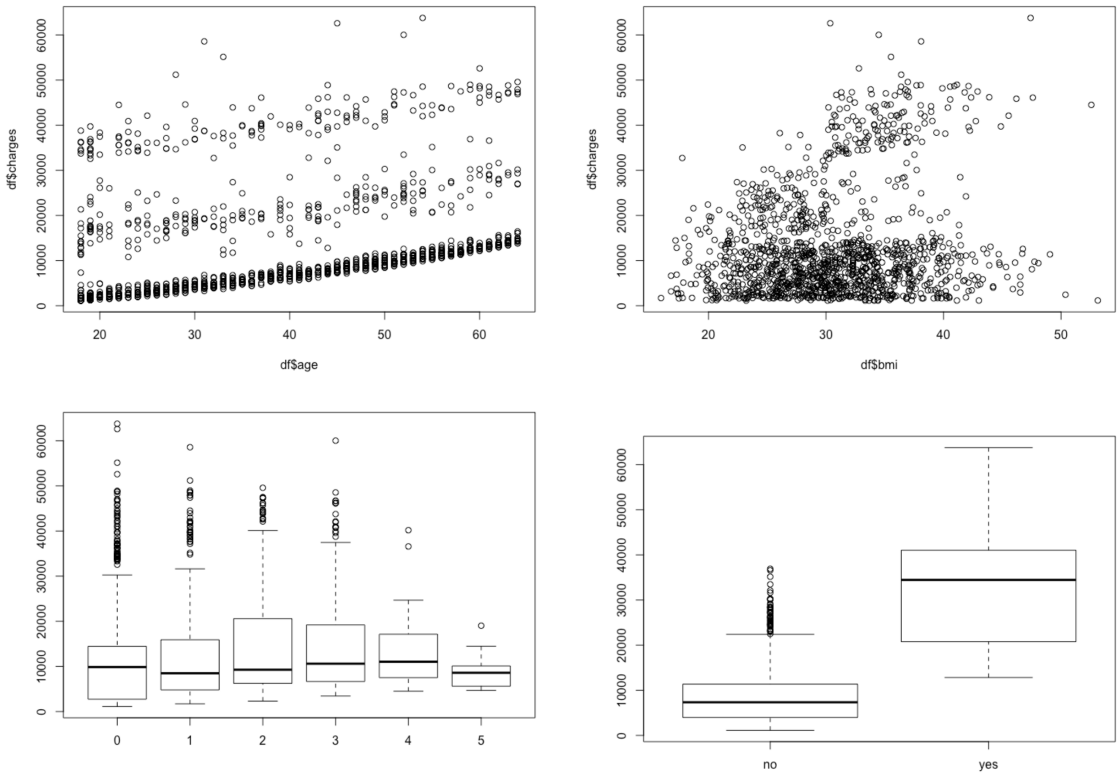
Usporedbom tog modela s modelom proširenim varijablom 'regija', koristeći χ^2 test, vidimo da je mjesto stanovanja značajno te ga ne treba izbaciti iz modela:

```
Analysis of Variance Table

Model 1: charges ~ (age + sex + bmi + children + smoker + region)
                   - region + smoker:bmi
Model 2: charges ~ age + sex + bmi + children + smoker + region
                   + smoker:bmi
  Res.Df        RSS Df  Sum of Sq Pr(>Chi)
1   1328 3.1192e+10
2   1331 3.1519e+10  3  326682767 0.003032 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Promotrimo sada model koji ne sadrži spol osobe:

```
fit4 <- lm(charges~. + smoker:bmi - sex, data=df)

Analysis of Variance Table

Model 1: charges ~ age + sex + bmi + children + smoker + region
                   + smoker:bmi - sex
Model 2: charges ~ age + sex + bmi + children + smoker + region
                   + smoker:bmi
  Res.Df        RSS Df Sum of Sq Pr(>Chi)
1   1328 3.1192e+10
2   1329 3.1275e+10 11 82715239  0.06057 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kao što možemo vidjeti, spol osobe ne izgleda značajan na razini 95% za predviđanje cijene premije te stoga ćemo varijablu 'spol' osiguranika isključiti iz daljnje analize.



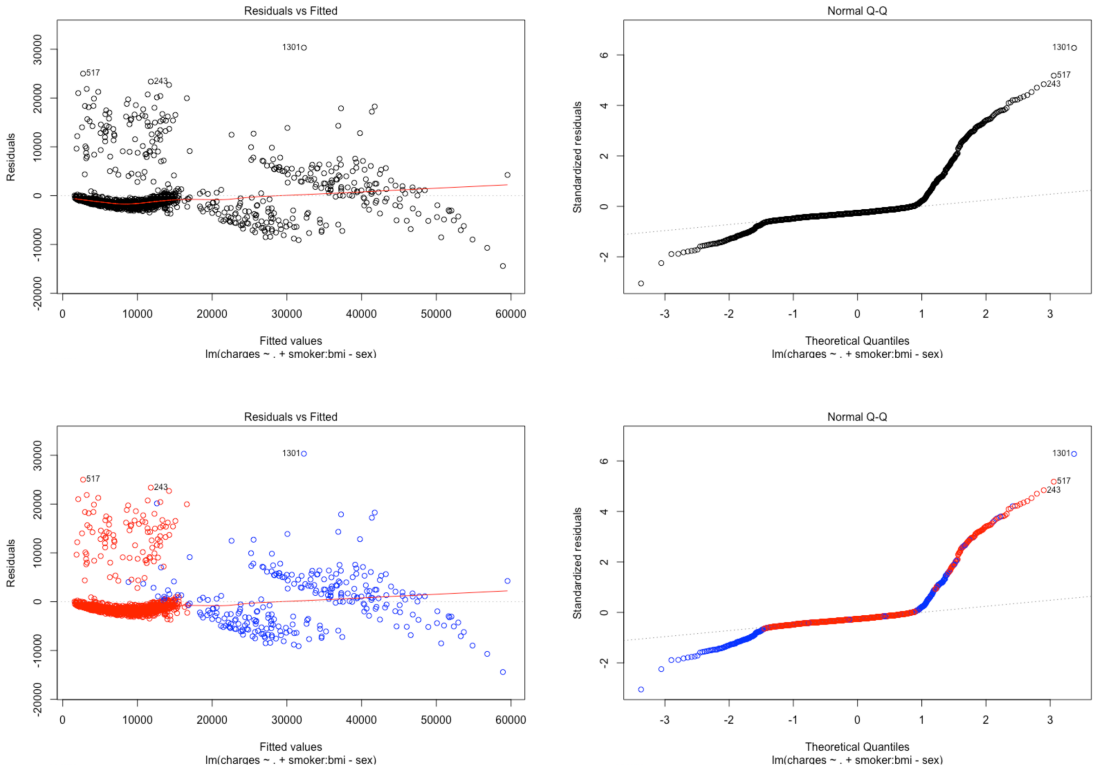
Slika 7.3: Prikaz ovisnosti medicinskih troškova o dobi klijenata (gore lijevo), njihovom indeksu tjelesne mase (gore desno), broju djece (dolje lijevo) te da li je klijent pušač (dolje desno).

Iz grafova ovisnosti medicinskih troškova o raznima varijablama: 'dob', 'indeks tjelesne mase', 'pušač' i 'broj djece', prikazanih na slici 7.1, možemo uistinu i primijetiti odokativno da trošak ovisi o svim tim varijablama. No, valja primijetiti da broj djece ne utječe monotono na iznos troškova, nego trošak jako varira u ovisnosti o broju djece. Ako, na primjer, osiguranik ima dvoje djece, to ne znači nužno da je dodatni utjecaj na cijenu premije dva puta veći nego što bi bio kod osiguranika koji ima samo jedno dijete. Stoga ćemo pretvoriti varijablu 'djeca' u kategoričku te model koji prilagođavamo ima po jedan koeficijent za svaki broj djece:

```
df$children = as.factor(df$children)
fit5 <- lm(charges~ . + smoker:bmi - sex, data=df)

Call:
lm(formula = charges ~ . + smoker:bmi - sex, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-14409.4  -1935.8  -1252.4   -370.2  30315.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2415.694    861.889  -2.803 0.005140 **
age                264.674      9.528  27.778  < 2e-16 ***
bmi                 20.789     25.604   0.812 0.416966
children1          315.393    336.908   0.936 0.349373
children2         1527.444    373.190   4.093 4.52e-05 ***
children3         1004.092    438.277   2.291 0.022120 *
children4         3397.226    990.867   3.429 0.000625 ***
children5         1805.257   1164.405   1.550 0.121292
smokeryes       -20292.082   1647.153 -12.319  < 2e-16 ***
regionnorthwest   -595.516    381.158  -1.562 0.118436
regionsoutheast  -1203.015    383.189  -3.139 0.001730 **
regionsouthwest  -1225.133    382.480  -3.203 0.001392 **
bmi:smokeryes     1437.623     52.586  27.338  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4845 on 1325 degrees of freedom
Multiple R-squared:  0.8414,Adjusted R-squared:  0.8399
F-statistic: 585.7 on 12 and 1325 DF,  p-value: < 2.2e-16
```

Na sličan način kao i prije možemo χ^2 testom testirati značajnost varijable 'djeca', te dobivamo da je ona uistinu i značajna sa p-vrijednosti $2. \times 10^{-5}$, bez obzira što neki koeficijenti nisu značajni. Valja napomenuti da kada se koriste kategoričke varijable nema nikakvog smisla odbacivati samo neke razine te kategoričke varijable.

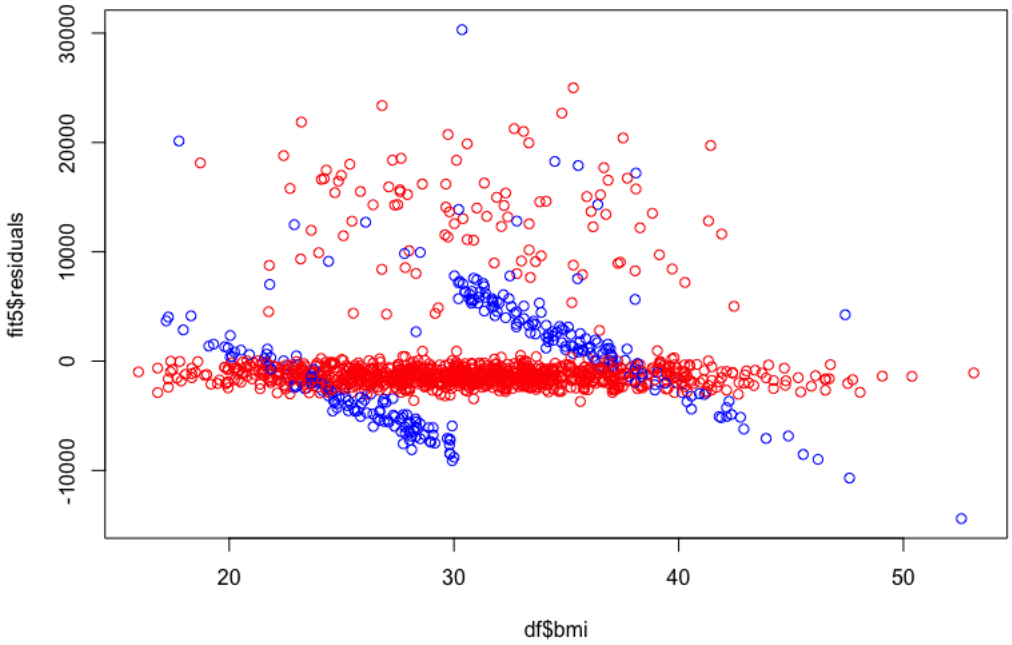
Da bi procijenili kvalitetu prilagodbe ovog modela, promotrimo dijagnostičke dijagrame. Na slici 7.4 gore lijevo vidimo Tukey-Anscombe graf, koji je, iako daleko od odličnog, mnogo bolji nego što je bio za linearni model prilagođen koristeći sve podatke, prikazan na slici 7.1.



Slika 7.4: Tukey-Anscombe graf (gore lijevo) i QQ graf (gore desno) za krajnji model. Na donjem dijelu slike su prikazani ti isti grafovi, gdje su točke obojane crveno ako je osoba nepušač, a plavo ako je pušač.

QQ graf, prikazan na slici 7.4 gore desno, je mnogo lošiji te pokazuje veoma teške repove distribucije. Međutim, ako obojamo točke na dijagnostičkim grafovima, što možemo vidjeti na donjem dijelu slike 7.4, vidimo kako teški repovi odgovaraju pušačima te bi bilo razumno prilagoditi potpuno novi model za pušače koji bi bolje se prilagodio podacima.

Ako promotrimo graf na slici 7.5, koji prikazuje reziduale krajnje regresije u ovisnosti o indeksu tjelesne težine, pri čemu su pušači obojani plavo, a nepušači crveno, možemo uočiti uzorak. Vidimo da za pušače greška prilagodbe modela ovisi izrazito nelinearno o indeksu tjelesne mase. Stoga bi bilo pametno pokušati prilagoditi modele koje dozvoljavaju nelinearan utjecaj interakcije između varijabli 'indeks tjelesne mase' i 'pušač'.



Slika 7.5: Prikaz ovisnosti reziduala u krajnjem modelu o indeksu tjelesne mase klijenta, pri čemu plave točke odgovaraju, a crvene odgovaraju nepušačima.

Kao što smo vidjeli, linearni model nije idealan za opisivanje veze među danim podacima, ali ipak model nije niti loš. Vidimo da model objašnjava čak 84% rasipanja podataka, što je veoma dobro. Linearni model, iako poprilično svestran, ipak često zna biti ograničen. Međutim, kao što smo mogli i vidjeti, linearni modeli omogućuju lakše istraživanje podataka te su odlična polazna točka za prilagodbu kompleksnijih modela.

Bibliografija

[1] R. V. Hogg, J. McKean, and A. T. Craig, *Introduction to mathematical statistics*. Pearson Education, 2005.

[2] E. W. Frees, *Regression modeling with actuarial and financial applications*. Cambridge University Press, 2009.

[3] A. L. Bailey, "A generalized theory of credibility," in *Proceedings of the Casualty Actuarial Society*, vol. 32, pp. 13–20, 1945.

[4] H. Bühlmann and A. Gisler, *A course in credibility theory and its applications*. Springer Science & Business Media, 2006.

[5] J. Lemaire, *Bonus-malus systems in automobile insurance*, vol. 19. Springer science & business media, 2012.

[6] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li, *et al.*, *Applied linear statistical models*, vol. 5. McGraw-Hill Irwin Boston, 2005.

[7] M. J. Crawley, *The R book*. John Wiley & Sons, 2012.

Sažetak

U ovom radu proučili smo primjene linearne regresije u aktuarstvu. Prvi dio obuhvaća teoriju potrebnu za razumijevanje sadržaja: formu i prilagodbu modela, nužne uvjete za valjanost modela te metode testiranja relevantnih svojstava. Dodatno smo promotrili dijagnostičke grafove koji su specifični za linearnu regresiju i koji su od velike pomoći pri detaljnijoj analizi prilagodbe modela. Na osnovu dva primjera predviđanja cijene premije osiguranja smo imali priliku vidjeti linearnu regresiju na djelu. Iako linearna regresija nije idealna za dane podatke, nije niti u potpunosti loša. Iz dobivene analize nam je jasno u kojem smjeru bismo dalje trebali ići, tako da je postavljena dobra osnova za daljni razvoj modela.

Summary

In this thesis, we have investigated application of linear regression in actuarial science. First part covers the theory necessary for understanding: model description, model fitting methods, model assumptions and methods for testing certain relevant properties. We have additionally considered diagnostic plots specific to linear regression, which are of great help with detailed analysis of the goodness-of-fit. We have investigated the performance of the linear regression based on two examples of the insurance premium prediction. Even though the linear regression is not ideal for the given datasets, it is not very bad either. From the analysis one can see how to proceed and thus the linear model is a good starting point for more complicated models.

Životopis

Rođena sam 21. lipnja 1995. godine u Zagrebu. Završila sam osnovnu školu Ljube Babića u Jastrebarskom 2010. godine i 2010./2011. upisala sam XV. gimnaziju u Zagrebu koju sam završila 2014. godine. U akademskoj godini 2014./2015. upisujem preddiplomski sveučilišni studij matematike na matematičkom odsjeku PMF-a u Zagrebu. Preddiplomski studij završavam 2017. godine te u akademskoj godini 2017./2018. upisujem diplomski studij, smjer Financijska i poslovna matematika na matematičkom odsjeku PMF-a u Zagrebu, gdje ga u jesen 2019. godine i završavam.