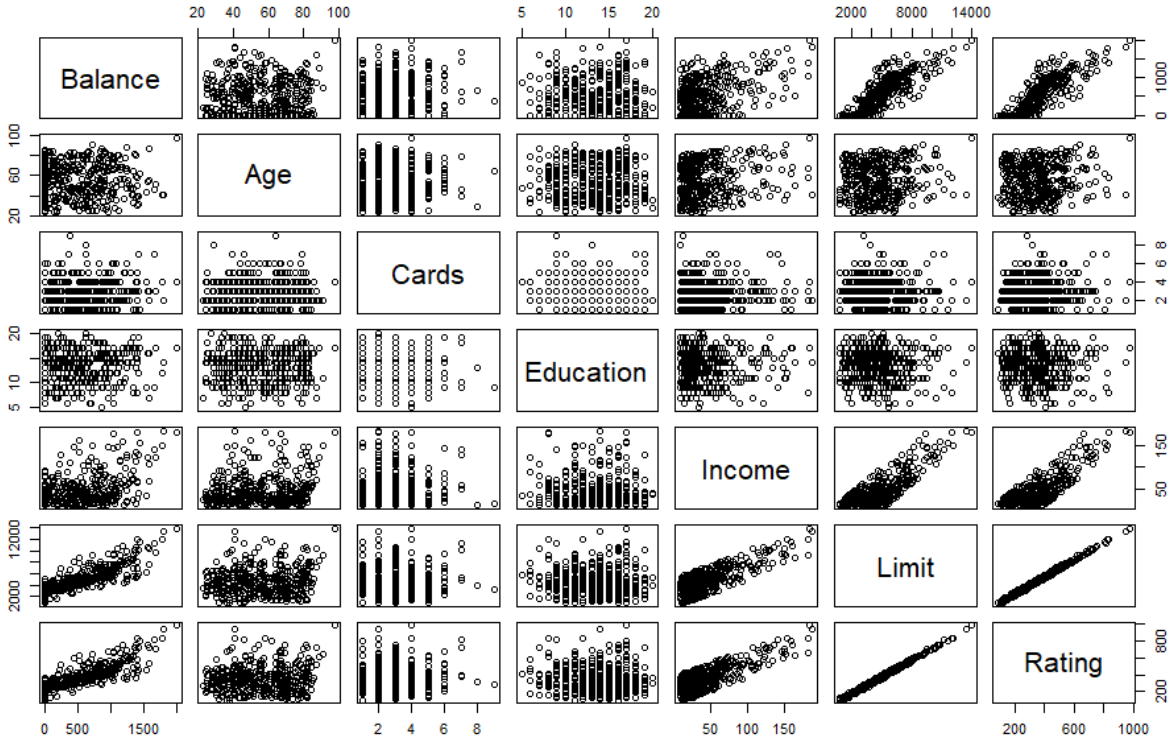
Još po nešto o regresionim modelima

Kvalitativni prediktori

Prikaz za *Credit* skup podataka iz *ISLR* biblioteke.

U R-u: `pairs(Credit)`



Kvalitativni prediktori sa samo dva nivoa

$$x_i = \begin{cases} 1, & \text{ako je } i-\text{ta osoba ženskog pola} \\ 0, & \text{ako je } i-\text{ta osoba muškog pola} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + e_i = \begin{cases} \beta_0 + \beta_1 + e_i, & \text{ako je } i-\text{ta osoba ženskog pola} \\ \beta_0 + e_i, & \text{ako je } i-\text{ta osoba muškog pola} \end{cases}$$

Sada se β_0 može tumačiti kao prosečno stanje na kreditnoj kartici kod muškaraca, $\beta_0 + \beta_1$ kao prosečno stanje na kreditnim karticama među osobama ženskog pola, a β_1 kao prosečna razlika u stanju na kreditnim karticama između osoba ženskog i muškog pola.

*****

Alternativno, umesto šeme kodiranja 0/1, mogli bismo da kreiramo *dummy* promenljivu na sledeći način:

$$x_i = \begin{cases} 1, & \text{ako je } i-\text{ta osoba ženskog pola} \\ -1, & \text{ako je } i-\text{ta osoba muškog pola} \end{cases}$$

Sada se β_0 može protumačiti kao ukupno prosečno stanje na kreditnoj kartici (zanemarujući efekat pola), a β_1 je iznos za koji su žene iznad, a muškarci ispod proseka.

Važno je napomenuti da će konačna predviđanja za kreditne bilance muškaraca i žena biti identična bez obzira na šemu kodiranja koja se koristi. Razlika je samo u načinu na koji se koeficijenti tumače.

Kvalitativni prediktori sa više nivoa

Kada kvalitativni prediktor ima više od dva nivoa, tada jedna *dummy* promenljiva ne može predstavljati sve moguće vrednosti. U ovoj situaciji možemo kreirati dodatne *dummy* promenljive. Ako, na primer, imamo kategoriju sa 3 nivoa (Azijat, Kavkaz i Afro-amerikanac), tada:

$$x_{i1} = \begin{cases} 1, & \text{ako je } i-\text{ta osoba Azijat} \\ 0, & \text{ako je } i-\text{ta osoba nije Azijat} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{ako je } i-\text{ta osoba sa Kavkaza} \\ 0, & \text{ako je } i-\text{ta osoba nije sa Kavkaza} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i = \begin{cases} \beta_0 + \beta_1 + e_i, & \text{ako je } i-\text{ta osoba Azijat} \\ \beta_0 + \beta_2 + e_i, & \text{ako je } i-\text{ta osoba sa Kavkaza} \\ \beta_0 + e_i, & \text{ako je } i-\text{ta osoba Afro}-\text{amerikanac} \end{cases}$$

Uvek će postojati jedna promenljiva manje od broja postojećih nivoa. Nivo bez *dummy* promenljive – Afro-američki u ovom primeru - poznat je i kao *osnovni nivo* (eng. *baseline*).

Radna verzija

```
> model = lm(Balance ~ Ethnicity, data = Credit)
> summary(model)

Call:
lm(formula = Balance ~ Ethnicity, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         531.00      46.32  11.464   <2e-16 ***
EthnicityAsian      -18.69      65.02  -0.287    0.774
EthnicityCaucasian  -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

Tabela 1. Predikcija stanja na kartici u zavisnosti od etničke pripadnosti.

Interpretacija

Iz Tabele 1 vidimo da je procenjeno stanje za osnovnu liniju, Afro-američka, iznosi $531. Procenjuje se da će azijska kategorija imati 18.69 dolara niže stanje na računu od afroameričke kategorije, a kavkaska 12.50 dolara manje od afroameričke kategorije. Međutim, p-vrednosti povezane sa procenama koeficijenta za ove dve *dummy* promenljive su veoma velike, što ne ukazuje na statističke dokaze o stvarnoj razlici u stanju na kreditnoj kartici između etničkih grupa. Još jednom, nivo izabran kao osnovna kategorija je proizvoljan i konačna predviđanja za svaku grupu biće ista bez obzira na ovaj izbor. Međutim, koeficijenti i njihove p-vrednosti zavise od izbora dummy kodiranja promenljive. Umesto da se oslanjamo na pojedinačne koeficijente, možemo da koristimo *F*-test za ispitivanje hipoteze H_0: $\beta_1 = \beta_2 = 0$; F-test ne zavisi od načina kodiranja. F-test ima p-vrednost 0.96, što ukazuje na to da ne možemo odbaciti nultu hipotezu da ne postoji veza između stanja na računu i etničke pripadnosti (ne možemo sa sigurnošću tvrditi da postoji veza).

Pored toga postoji mnogo različitih načina za kodiranja kvalitativnih promenljivih pored ovde korišćenog pristupa sa *dummy* promenljivama. Svi ovi pristupi dovode do ekvivalentnih modela, ali koeficijenti su različiti, imaju različita tumačenja i dizajnirani su za merenje određenih kontrasta.

Radna verzija

Proširenje linearnih modela

Standardni model linearne regresije (1) daje interpretabilne rezultate i prilično dobro radi na mnogim stvarnim problemima. Međutim, ovaj model donosi i nekoliko visoko restriktivnih pretpostavki koja su često prekršena u praksi.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e \quad (1)$$

Dve najvažnije pretpostavke navode da je veza između prediktora i odgovora **aditivna (eng. *additive*) i linearna (eng. *linear*)**. Pretpostavka aditiva znači da je efekat promena prediktora X_j na odgovor Y nezavisan od vrednosti koje imaju ostali prediktori. Linearna pretpostavka navodi da je promena outputa Y usled promene X_j za jednu jedinicu konstantna, bez obzira na vrednost X_j. U nastavku ispitujemo neke sofisticirane metode koje relaksiraju ove dve pretpostavke. Ovde ćemo ukratko ispitati neke uobičajene klasične pristupe proširenja linearnog modela.

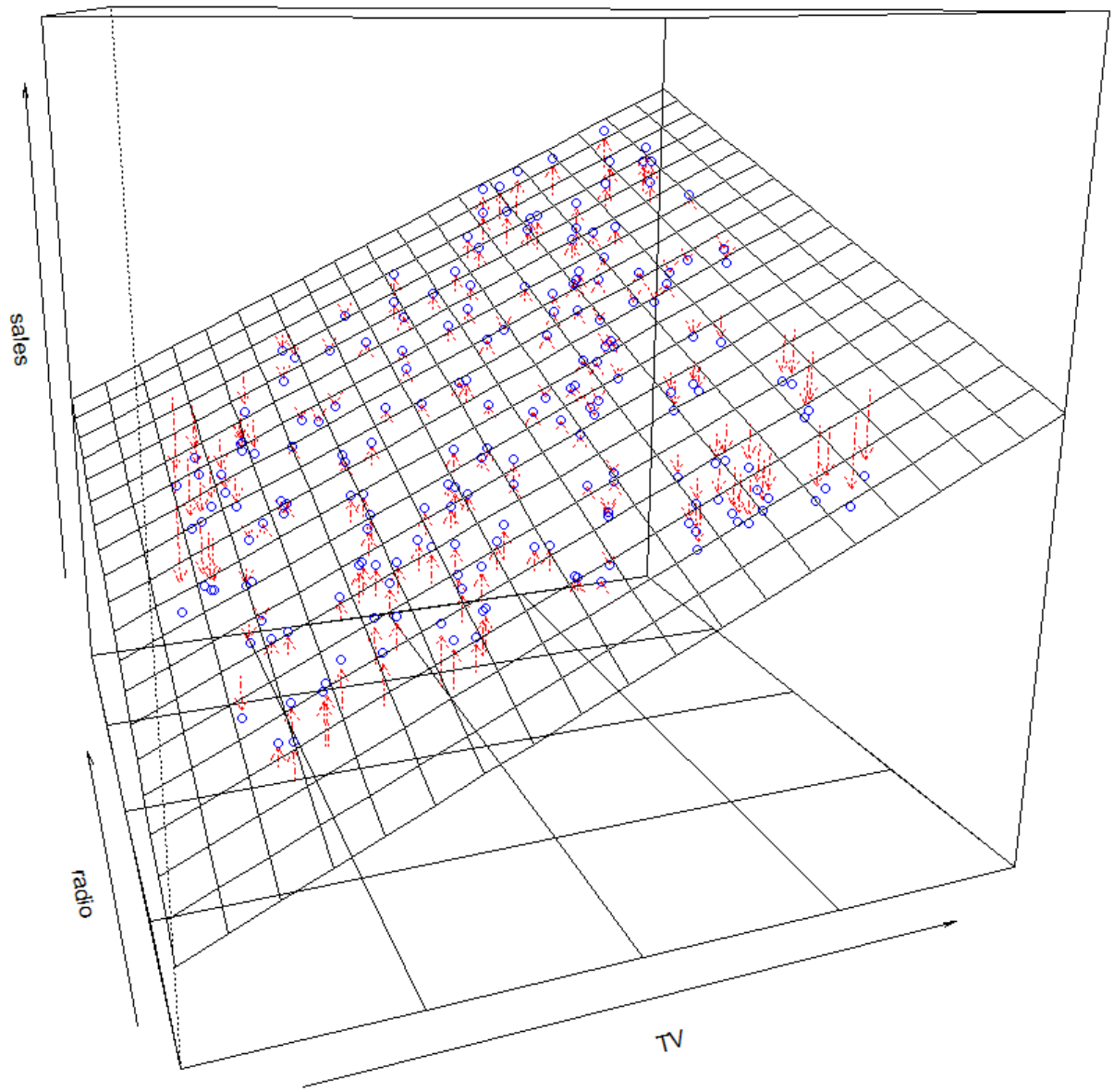
Uklanjanje aditivne pretpostavke

U prethodnoj analizi podataka o oglašavanju zaključili smo da se čini da su i TV i radio povezani sa prodajom. Linearni modeli koji su činili osnovu za ovaj zaključak pretpostavljali su da je efekat povećanja jednog reklamnog medija na prodaju nezavisan od količine novca potrošene na druge medije. *Na primer, jedan od prethodno prikazanih linearnih modela navodi da je prosečni efekat na prodaju porasta TV od jedne jedinice uvek β_1, bez obzira na količinu koja je potrošena na radio reklamiranje.*

Međutim, ovaj jednostavan model može biti netačan. Pretpostavimo da trošenje novac za radio oglašavanje zapravo povećava efikasnost TV oglašavanja, tako da bi trebalo da se nagib za TV povećava kako se radio povećava. U ovoj situaciji, s obzirom na fiksni budžet od $100 000, potrošnja polovine na radio, a polovine na TV-u može povećati prodaju više nego dodeljivanje celokupnog iznosa bilo TV-u ili radiju. U marketingu je ovo poznato kao *sinergijski (eng. sinergy) efekat*, a u statistici se to naziva efektom interakcije. Slika 1 sugeriše da takav efekat može biti prisutan u podacima o oglašavanju.

```
library(rockchalk)
model_tv_radio = lm(sales ~ TV + radio, data = advertising)
summary(model_tv_radio)
plotPlane(model_tv_radio, plotx1 = "TV", plotx2 = "radio", drawArrows = TRUE)
```

Radna verzija



Slika 1. Iz obrasca ostataka to možemo videti u podacima postoji izražena nelinearna veza. Pozitivni ostaci (oni vidljivi iznad površine) imaju tendenciju da leže duž linije od 45 stepeni, gde su budžeti za TV i Radio ravnomerno podeljeni. Negativni rezidual (većina nije vidljiv) imaju tendenciju da se udaljavaju od ove linije, gde su budžeti više neodređeni.

Razmotrimo standardni model linearne regresije sa dve promenljive,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e.$$

Prema ovom modelu, ako X_1 povećamo za jednu jedinicu, tada će Y porasti za prosečno β_1. Primetite da prisustvo X_2 ne menja ovu izjavu - to jest, bez obzira na vrednost X_2, povećanje X_1 za jednu jedinicu će dovesti do povećanja β_1-jedinice u Y. Jedan od načina za proširivanje ovog

režima kako bi se omogućili efekti interakcije je uključivanje trećeg prediktora, koji se naziva interakcijom, a koji je konstruisan izračunavanjem proizvoda X_1 i X_2. Ovo rezultira modelom:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e. \quad (2)$$

Kako uključivanje ovog pojma interakcije relaksira aditivnu pretpostavku? Primetite da se (2) može zapisati kao

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + e \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + e \end{aligned}$$

gde je $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$. Pošto se $\tilde{\beta}_1$ menja sa X_2, efekat X_1 na Y više nije konstanta: menjanje X_2 će promeniti i uticaj X_1 na Y.

***** R kod za sinergijski efekat

```
model_sinergy = lm(sales ~ TV * radio, data = advertising)
summary(model_sinergy)
```

*****

Interpretacija

PRIMER 1: Na primer, pretpostavimo da smo zainteresovani za proučavanje nivoa produktivnosti u okviru neke fabrike. Želimo da predvidimo broj proizvedenih jedinica na osnovu broja proizvodnih linija i ukupnog broja radnika. Jasno je da će efekat povećanja broja proizvodnih linija zavisiti od broja radnika, jer ako nema raspoloživih radnika za upravljanje linijama, povećanje broja linija neće povećati proizvodnju.

$$\begin{aligned} units &\approx 1.2 + 3.4 \times lines + 0.22 \times workers + 1.4 \times (lines \times workers) \\ &= 1.2 + (3.4 + 1.4 \times workers) \times lines + 0.22 \times workers \end{aligned}$$

Drugim rečima, dodavanjem dodatne linije povećaće se broj proizvedenih jedinica za $3{,}4 + 1{,}4 \times workers$. Otuda što više radnika imamo, snažniji će biti efekat proizvodnih linija.

PRIMER 2: Primer vezan za oglašavanje. Linearni model koji koristi radio, TV i interakciju između njih dva prediktora za predviđanje prodaje ima oblik:

$$\begin{aligned} sales &= \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \\ &= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + e. \end{aligned} \quad (3)$$

Možemo tumačiti β_3 kao povećanje efikasnosti TV oglašavanja za jednu jedinicu povećanja radio oglašavanja (ili obrnuto).

Radna verzija

Rezultati u modelu snažno sugerišu da je model koji uključuje pojam interakcije superiorniji od modela koji sadrži samo glavne efekte.

R^2 za model (3) je 96.8%, u poređenju sa samo 89.7% za model koji predviđa prodaju pomoću TV-a i radija bez interakcije. To znači da je (96.8 – 89.7) / (100 −89.7) = 69% varijabilnosti prodaje koja ostaje nakon uklapanja aditivnog modela objašnjeno terminom interakcije.

Dobijeni koeficijenti (**napravite sami model!**) sugerišu da je povećanje TV oglašavanja od \$1000 povezano sa povećanom prodajom $\left(\hat{\beta}_1 + \hat{\beta}_3 \times radio\right) \times 1000 = 19 + 1.1 \times radio$ jedinica. A povećanje radio oglašavanja od \$1.000 biće povezano sa povećanjem prodaje od $\left(\hat{\beta}_2 + \hat{\beta}_3 \times TV\right) \times 1000 = 29 + 1.1 \times TV$ jedinica.

Može se ponekad desiti da pojam interakcije ima vrlo malu p-vrednost, ali povezani glavni efekti (u ovom slučaju TV i radio) nemaju. **Hijerarhijski princip** kaže da *ako u model uključimo interakciju, tada bi trebalo da uključimo i glavne efekte, čak iako p-vrednosti povezane sa njihovim koeficijentima nisu značajne.* Takođe, $X_1 \times X_2$ je u korelaciji sa X_1 i X_2, pa tako njihovo izostavljanje teži da promeni smisao interakcije.

*****

Interakcija između kvalitativne promenljive i kvantitativne promenljive ima posebno lepo tumačenje.

Ako posmatramo podatake o kreditima (*Credit*) i pretpostavimo da želimo da predvidimo stanje na računu koristeći promenljive prihod (kvantitativne) i da li je osoba student (kvalitativne). U odsustvu termina interakcije, model ima oblik:

$$balance_i = \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2, & \text{ako je } i-\text{ta osoba student} \\ 0, & \text{ako je } i-\text{ta osoba nije student} \end{cases}$$

$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2, & \text{ako je } i-\text{ta osoba student} \\ \beta_0, & \text{ako je } i-\text{ta osoba nije student} \end{cases}$$

Primetite da ovi modeli predstavljaju dve paralelne linije koje predstavljaju podatke, jednu za studente i jednu za osobe koje nisu studenti. Linije za osobe koje su studenti i koje nisu studenti imaju različite preseke, $\beta_0 + \beta_2$ nasuprot β_0, ali isti nagib, β_1. Činjenica da su linije paralelne znači da prosečni efekat povećanja stanja na računu u zavisnosti od jediničnog povećanja prihoda ne zavisi od toga da li je pojedinac student ili ne. Ovo predstavlja potencijalno ozbiljno ograničenje modela, jer promena prihoda može imati vrlo različit efekat na stanje na kreditnoj kartici studenta u odnosu na one osobe koje nisu studenti.

Ovo ograničenje se može rešiti dodavanjem promenljive za interakciju, stvorene množenjem prihoda sa *dummy* promenljivom za studenta.

Radna verzija

Sada naš model postaje:

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 + \beta_3 \times income_i, & \text{ako je } i-\text{ta osoba student} \\ 0, & \text{ako je } i-\text{ta osoba nije student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income_i, & \text{ako je } i-\text{ta osoba student} \\ \beta_0 + \beta_1 \times income_i, & \text{ako je } i-\text{ta osoba nije student} \end{cases}$$

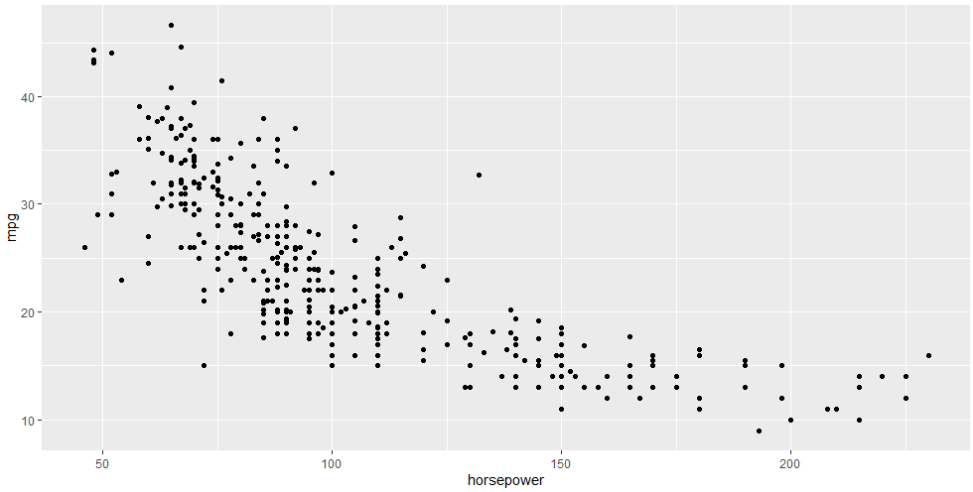
Sada imamo dve različite regresione linije za studente i za one koji nisu studenti. Ali sada te regresione linije imaju različite preseke, $\beta_0 + \beta_2$ nasuprot β_0, kao i različite nagibe, $\beta_1 + \beta_3$ nasuprot β_1. Ovo omogućava mogućnost da promene prihoda mogu različito uticati na stanje na kreditnim karticama studenata i ne-studenata. Primećujemo da je nagib za studente niži od nagiba za ne-studente (**Proveriti ovo tvrđenje!**). To sugeriše da su **povećanja prihoda povezana sa manjim povećanjem stanja na kreditnim karticama** među studentima u poređenju sa ne-studentima.

Nelinearne veze

Kao što je prethodno razmatrano, model linearne regresije (1) pretpostavlja linearni odnos između outputa i prediktora. Ali u nekim slučajevima pravi odnos između outputa i prediktora može biti nelinearan. Ovde predstavljamo vrlo jednostavan način da se linearni model direktno proširi kako bi se prilagodio nelinearnim odnosima, koristeći polinomsku regresiju.

Biblioteka: **ISLR**

```
ggplot(data = Auto) + geom_point(aes(x = horsepower, y = mpg))
```



Radna verzija

Jednostavan pristup za uključivanje nelinearnih asocijacija u linearni model je uključivanje transformisanih verzija prediktora u model. Tačke na Slici ## imaju „kvadratni“ oblik, što sugeriše da model možemo poboljšati na sledeći način:

$$mpg = \beta_0 + \beta_1 \times \text{horspower} + \beta_2 \times \text{horspower}^2 + e. \quad (4)$$

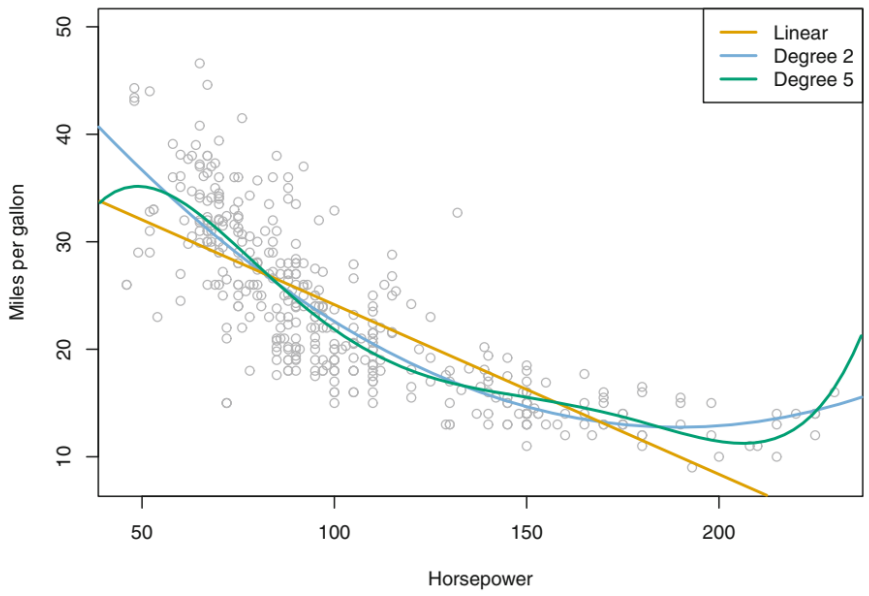
Jednačina (4) uključuje predviđanje *mpg* pomoću nelinearne funkcije u zavisnosti od konjskih snaga. **I dalje imamo linearni model!** To je i dalje model višestruke linearne regresije. Na ovaj način postižemo nelinearno modelovanje podataka.

*** R kod

```
model_auto2 = lm(mpg ~ poly(horsepower, 2), data = Auto)
summary(model_auto2)
predicted.intervals = predict(model_auto2, x=Auto$horsepower,
                                          interval='confidence', level = 0.99)
g = ggplot(data = Auto) + geom_point(aes(x = horsepower, y = mpg))
g + geom_line(aes(y=predicted.intervals[, 1], x=Auto$horsepower), colour='blue')
```

***

Za domaci! Nacrtati grafik sa slike!



Pristup koji smo upravo opisali za proširenje linearnog modela kako bi se prilagodio nelinearnim vezama je poznat pod nazivom **polinomska regresija**, jer smo u regresioni model uključili polinomske funkcije prediktora.

Kako da znamo do kojeg stepena da dodajemo članove? Na nekom od narednih termina naučićemo šta je regularizacija i kako nam ona pomaže u davanju odgovora na ovo pitanje!!!

Probna verzija