Linearna regresija

Linearna regresija je vrlo jednostavan nadgledani pristup učenju. Linearna regresija je naročito koristan alat za predviđanje kvantitativnog odgovora.

Zadatak: Pretpostavimo da se u ulozi statističkih konsultanata od nas traži da na osnovu *Advertising* podataka predložimo marketinški plan za sledeću godinu koji će rezultirati velikom prodajom proizvoda.

Evo nekoliko važnih pitanja koja bi trebalo da damo odgovor:

1. **Da li postoji veza između budžeta za oglašavanje i prodaje?**

   Prvi cilj je da pružimo dokaz o povezanosti između troškova oglašavanja i prodaje. Ako su dokazi slabi, onda se može tvrditi da nikakav novac ne bi trebalo trošiti na oglašavanje!

2. **Koliko je jak odnos između budžeta za oglašavanje i prodaje?**
3. **Koji mediji doprinose prodaji?**

   Da bismo odgovorili na ovo pitanje, moramo pronaći način da razdvojimo pojedinačne efekte svakog medija kada smo potrošili novac na sva tri medija.

4. **Koliko tačno možemo proceniti efekat svakog medija na prodaju?**

   Za svaki dolar potrošen na oglašavanje u određenom mediju, za koji iznos će se povećati prodaja? Koliko tačno možemo ovo predvideti iznos povećanja?

5. **Koliko tačno možemo predvideti buduću prodaju?**
6. **Da li je odnos linearan?**
7. **Postoji li sinergija među reklamnim medijima?**

   Možda potrošiti 50.000 USD na televizijsko oglašavanje i 50.000 USD na radio oglašavanje rezultira većom prodajom od izdvajanja 100.000 američkih dolara bilo televizija ili radio pojedinačno. U marketingu je ovo poznato kao *efekat sinergije*, dok se u statistici naziva *efektom interakcije*.

Jednostavna linearna regresija opravdava svoje ime: to je vrlo neposredan, jednostavan linearni pristup za predviđanje kvantitativnog odgovora Y na osnovu jedne prediktorske promenljive X. Pretpostavlja se da postoji približno linearna veza između X i Y. Matematički ovaj linearni odnos možemo zapisati kao:

$$Y \approx \beta_0 + \beta_1 X.$$

Na primer, X može predstavljati TV oglašavanje, a Y prodaja. Tada možemo regresirati prodaju na TV oglašavanju postavljanjem modela:

$$sales \approx \beta_0 + \beta_1 \times TV.$$

U datoj jednačini, β_0 i β_1 su dve nepoznate konstante koje predstavljaju presek (eng. *intercept*) i nagiba (eng. *slope*) u linearnom modelu. Zajedno su β_0 i β_1 poznati kao koeficijenti ili parametri modela.

Nakon što smo koristili trening podatke za određivanje parametara modela, $\hat{\beta}_0$ i $\hat{\beta}_1$, može se predvideti buduća prodaja na osnovu određene vrednosti TV oglašavanja računanjem

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

gde $\hat{y}$ označava predviđanje za Y na osnovu $X = x$.

Određivanje koeficijenata

Podaci: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

Neka je $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ predikcija za Y na osnovu $X = x_i$. Tada, $e_i = y_i - \hat{y}_i$ predstavlja i-ti residual linearnog modela. Tada se definiše sledeća veličina *residual sum of squares (RSS)* kao:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2.$$

Cilj je minimizovati RSS vrednost. Pokazuje se da važi sledeće:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ove jednačine predstavljaju metodu najmanjih kvadrata (eng. *least squares coefficient estimates*) za koeficijente proste linearne regresije.

Primer: Advertising dataset

```
> model = lm(sales ~ ., data = adver)
> summary(model)

Call:
lm(formula = sales ~ ., data = adver)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Tabela 1. Sales ~ TV + Radio + Newspaper

Preciznost dobijenih koeficijenata

Prava veza između X i Y ima formu $Y = f(X) + e$ za neku nepoznatu funkciju f, a e je greška sa 0-prosekom. Ako f aproksimiramo linearnom funkcijom, tada se dobije veza u obliku:

$$Y = \beta_0 + \beta_1 X + e \qquad (1)$$

Ovde je β_0 *inception* - odnosno očekivana vrednost Y kada je $X = 0$, a β_1 nagib - prosečno povećanje Y povezano sa povećanjem X za *jednu jedinicu mere*.

Greška je sve ono što nam nedostaje u ovom jednostavnom modelu: pravi odnos verovatno nije linearan, možda postoji druge promenljive koje uzrokuju varijaciju X, a možda postoji i greška merenja. **Obično se pretpostavlja da je greška nezavisna od X.**

Model dat u (1) definiše regresiju za **populaciju**, koja predstavlja najbolju linearnu aproksimaciju pravog odnosa između X i Y [1]. Procenjeni koeficijenti regresije pomoću metode najmanjih kvadrata (prema ranije datim formulama) karakterišu liniju najmanjih kvadrata (eng. *least squares line*): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

[1] Pretpostavka da je model linearan je često koristan radni model. Međutim, uprkos onome što nam pričaju mnogi udžbenici, retko se veruje da je pravi odnos linearni.

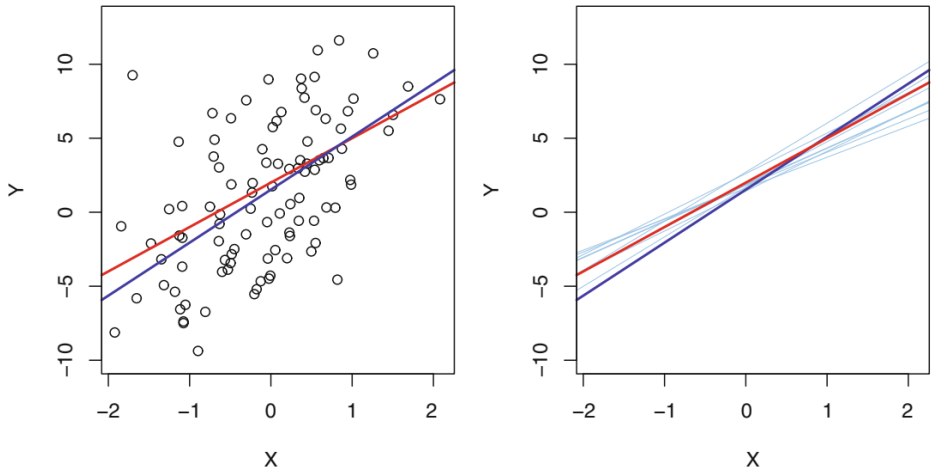
Populacija vs uzorak

Populacija je skup svih članova koji imaju određenu, zajedničku karakteristiku; skup svih ljudi ili stvari koji su od interesa u određenom istraživanju.

Veličina populacije je određena brojem svojih članova.

Uzorak je manji skup ljudi ili stvari koji su izabrani iz osnovnog skupa, tj. populacije. **Uzorak** je, dakle, **podskup** osnovnog skupa, odnosno **populacije**.

Osnovna namena uzorka je da se na osnovu rezultata merenja ili ispitivanja na njemu **zaključuje** o populaciji.



Slika 1. Simulirani skup podataka. **Levo**: Crvena linija predstavlja pravu vezu, $f(X) = 2 + 3X$, koja je poznata kao linija regresije populacije. Plava linija je linija najmanjih kvadrata; **Desno**: prikazano je deset linija najmanjih kvadrata, svaka izračunata na osnovu zasebnog slučajnog skupa posmatranja.

Na prvi pogled, razlika između regresione linije populacije i linije najmanjih kvadrata može izgledati zbunjujuće. *Imamo samo jedan skup podataka - šta onda znači da dve različite linije opisuju odnos između prediktora i odgovora?* U osnovi, koncept ove dve linije prirodno je proširenje standardnog statističkog pristupa koji koristi informacije iz uzorka za procenu karakteristika velike populacije.

Na primer, pretpostavimo da nas zanima srednja vrednosti populacije μ neke slučajne promenljive Y. μ je nepoznato, jer nemamo pristup svim observacijama. Šta dalje?

Na isti način, nepoznati koeficijenti β_0 i β_1 u linearnoj regresiji definišu liniju regresiju populacije. Te nepoznate koeficijente nastojimo da procenimo pomoću $\hat{\beta}_0$ i $\hat{\beta}_1$. Ove procene koeficijenta definišu liniju najmanjih kvadrata.

Analogija između linearne regresije i procene srednje vrednosti slučajne promenljive je odgovarajuća i zasnovana je na konceptu *bias*-a. Ako koristimo srednju vrednost uzorka $\hat{\mu}$ za procenu μ, ova procena je nepristrasna (*unbiased*), u smislu da u proseku očekujemo da je $\hat{\mu}$ jednako μ. Neke su vrednosti veće, a neke manje. Šta ovo tačno ovo znači?

Ako bismo mogli da odredimo prosek za ogroman broj procena za μ dobijen iz ogromnog broja skupova posmatranja, tada bi ovaj prosek bio tačno jednak μ.

Da li svojstvo nepristrasnosti (eng. *unbiasedness*) važi i za procene koeficijenata najmanjih kvadrata?

*****

Dalje nastavljamo analogiju sa μ. Pitanje koje se prirodno nameće je: koliko tačno srednja vrednost uzorka $\hat{\mu}$ aproksimira vrednost μ?

Ustanovili smo da će prosek $\hat{\mu}$ tokom mnogih skupova podataka biti vrlo blizu μ, ali da pojedinačna procena $\hat{\mu}$ može biti značajno manja ili veća od μ. Koliko će loša biti ta pojedinačna procena $\hat{\mu}$?

Generalno, na ovo pitanje se može odgovoriti izračunavanjem standardne greške od $\hat{\mu}$, koja se obeležava sa $SE(\hat{\mu})$. Za nju važi dobro poznata formula:

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}, \tag{2}$$

gde je σ standardna devijacija svake od realizacija y_i od Y. Grubo govoreći, standardna greška nam govori o prosečnom iznosu za koji se procena $\hat{\mu}$ razlikuje od stvarne vrednosti μ. Jednačina (2) takođe nam govori kako se ovo odstupanje smanjuje sa n - što više zapažanja imamo, to je manja standardna greška od $\hat{\mu}$.

Na sličan način može se doći do formula:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3}$$

gde je $\sigma^2 = Var(e)$. **Da bi ove formule bile strogo validne, moramo pretpostaviti da greške e_i za svaku observaciju imaju istu varijansu σ^2 i da međusobno nisu u korelaciji[2]**. Generalno, σ^2 nije poznato, ali se može proceniti na osnovu podataka. Ova procena je poznata kao **rezidualna standardna greška (eng. *residual standard error*)**, a data je formulom:

$$\sigma = RSE = \sqrt{RSS/(n-2)}.$$

opciono

[2] https://stats.oarc.ucla.edu/spss/seminars/introduction-to-factor-analysis/

Strogo govoreći, kada se procenjuje σ^2 na osnovu podataka, trebalo bi da se napiše $\widehat{SE}(\hat{\beta}_1)$ da bismo ukazali da je napravljena procena, ali zbog jednostavnosti notacije „šešir“ se ne piše.

Za izračunavanje intervala poverenja (eng. *confidence interval*) mogu se koristiti standardne greške (3). 95% interval poverenja je definisan kao opseg vrednosti takav da će sa 95% verovatnoće opseg sadržati pravu nepoznatu vrednost parametra. Za linearnu regresiju, interval pouzdanosti od 95% za β_1 ima oblik:

$$\hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1).$$

Odnosno, postoji približno 95% šanse da interval

$$[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$$

sadrži pravu vrednost za β_1. Na sličan način se definiše i interval poverenja za β_0.

```
> model = lm(sales ~ TV, data = adver)
> summary(model)

Call:
lm(formula = sales ~ TV, data = adver)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Interpretacija: Sada možemo zaključiti da će u odsustvu bilo kakvog oglašavanja prodaja u proseku biti negde između 6.130 i 7.935 jedinica. Dalje, za svako povećanje televizijskog oglašavanja od 1 000 dolara, biće prosečan porast prodaje između 42 i 53 jedinice.

*****

Standardne greške se takođe mogu koristiti za izvođenje statističkih testova nad koeficijentima. Najčešće korišćeni test za proveru hipoteze uključuje ispitivanje nulte hipoteze:

H_0: Ne postoji veza između X i Y ($\beta_1 = 0$).

Protiv alternativne hipoteze

$$H_a: \text{Između } X \text{ i } Y \text{ postoji određena veza } (\beta_1 \neq 0).$$

U praksi se računa t-statistika data sa:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \tag{4}$$

koja meri broj standardnih odstupanja za koje je $\hat{\beta}_1$ udaljeno od 0. Ako zaista ne postoji veza između X i Y, onda očekujemo da će (4) imati t-raspodelu sa $(n-2)$ stepena slobode. Dalje se jednostavno može izračunati verovatnoća da se dobije bilo koja vrednosti koja je jednaka ili veća od $|t|$, pod pretpostavkom da je $\beta_1 = 0$. Ovu verovatnoću nazivamo p-vrednost.

p-vrednost tumačimo na sledeći način: mala p-vrednost ukazuje na to da je malo verovatno da će se slučajno uočiti tako značajna povezanost između prediktora i odgovora, ako takva povezanost između prediktora i odgovora u stvarnosti ne postoji.

Dakle, ako vidimo malu p-vrednost, onda možemo zaključiti da postoji veza između prediktora i odgovora. Odbacujemo nultu hipotezu (eng. *reject the null hypothesis*) - odnosno izjavljujemo da postoji veza između X i Y - ako je p-vrednost dovoljno mala. Tipične granične vrednosti za p-vrednost za odbacivanje nulte hipoteze su 5 ili 1%.

Procena tačnosti modela

Kvalitet linearne regresione se najčešće procenjuje koristeći dve povezane veličine: **residual standard error (RSE)** i $\boldsymbol{R^2}$ **statistiku**.

Residual Standard Error

Podsetimo se iz definicije modela da je svaka observacija povezana sa greškom e. Zbog prisustva ovih grešaka, čak i kada bismo znali pravu regresionu liniju (tj. čak i kad bi bili poznati β_0 i β_1), ne bismo mogli da savršeno predvidimo Y iz X. *RSE* je procena standardne devijacije od e. To je zapravo prosečan iznos za koji će output da odstupa od tačne regresione linije.

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

RSS (residual sum of squares) se definiše kao:

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

```
> model = lm(sales ~ TV, data = adver)
> summary(model)

Call:
lm(formula = sales ~ TV, data = adver)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

U slučaju podataka o oglašavanju (*advertisement*), stvarna prodaja u svakoj radnji u proseku odstupa od prave regresione linije za približno 3.260 jedinica. Drugi način razmišljanja o ovome je da čak i da je model tačan i da su tačno poznate prave vrednosti nepoznatih koeficijenata β_0 i β_1, svako predviđanje prodaje na osnovu **TV oglašavanja** i dalje bi se u proseku promašivalo za oko 3.260 jedinica.

Srednja vrednost prodaje na svim tržištima je približno 14 jedinica, tako da je procentualna greška 3.260 / 14 = 23%.

RSE se smatra merom neprilagođenosti modela podacima. Ako su predviđanja dobijena korišćenjem modela vrlo blizu istinitih vrednosti ishoda, tada će to biti RSE vrednost biti veoma mala i možemo zaključiti da model vrlo dobro odgovara podacima, i obrnuto.

R^2 Statistika

Statistika R^2 pruža alternativnu mera koliko naš model dobro fituje podatke. Formula za R^2 je:

$$R^2 = \frac{TSS - RSS}{TSS},$$

gde je $TSS = \sum(y_i - \bar{y})^2$ je ukupna suma kvadrata, a RSS je definisano ranije. $(TSS - RSS)$ meri količinu varijabilnosti u outputu koja je objašnjena izvođenjem regresije, a R^2 meri **procenat varijabilnosti u Y koji se može objasniti pomoću X**.

R^2 statistika je mera linearnog odnosa između X i Y. Za **prostu linearnu regresiju** važi da je $R^2 = Cor(X, Y)^2$. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Ovo ne važi u slučaju *multivarijantne linearne regresije*!

Potencijalni problemi sa R^2 metrikom:

Problem 1: Svaki put kada modelu dodate prediktor, R^2 se povećava. Nikad se ne smanjuje. Shodno tome, model sa više prediktora može izgledati kao bolja opcija samo zato što ima više prediktora.

Problem 2: Ako model ima previše prediktora i polinoma višeg reda, tada model počinje da modelira slučajni šum u podacima. Ovo stanje je poznato kao *overfitting*.

Prilagođeno/ponderisano R^2 ($adj\ R^2$) je modifikacija R^2 koja prilagođava broj prediktora p u modelu u odnosu na broj dostupnih observacija n. Prilagođeni R^2 je definisano kao

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

Multivarijantna linearna regresija

Jednostavna linearna regresija koristan je pristup za predviđanje odgovora na osnovu jedne prediktorske promenljive. Međutim, u praksi često imamo više prediktora. Na primer, u podacima o oglašavanju ispitali smo vezu između prodaje i TV oglašavanja. Kako možemo proširiti našu analizu podataka o oglašavanju redom da se prilagode ova dva dodatna prediktora?

Jedna od mogućnosti je pokretanje tri odvojene jednostavne linearne regresije, svaka od njih koja kao prediktor koristi drugačiji medij oglašavanja.

Neka imamo p različitih prediktora. Tada važi:

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

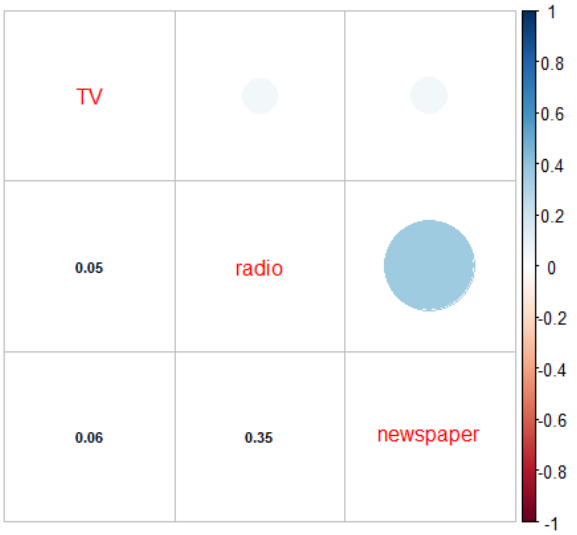
Tada, β_j interpretiramo kao prosečni efekat na Y kada se X_j poveća za jednu jedinicu mere, držeći sve ostale prediktore fiksnim.

Procena regresionih koeficijenata

Cilj je odrediti koeficijente $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ tako da suma kvadrata reziduala, RSS, bude minimalna. Vrednosti $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ koji minimiziraju RSS procene koeficijenta regresije višestrukih najmanjih kvadrata.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
```

- Interpretacija
- Matrica korelacija



Neka bitna pitanja

1. Da li je bar jedan od prediktora $X_1, X_2, \dots, X_p$ koristan u predviđanju odgovora?
2. Da li svi prediktori pomažu u objašnjavanju Y ili je samo podskup prediktora koristan?
3. Koliko dobro se model fituje sa podacima?
4. S obzirom na skup vrednosti prediktora, koju vrednost outputa bi trebalo da predvidimo, i koliko je tačno naše predviđanje?

1. Da li postoji veza između odgovora i prediktora?

Testira se nulta hipoteza:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{5}$$

Alternativna hipoteza je:

$$H_a: \text{najmanje jedno } \beta_j \text{ je različito od } 0.$$

Hipoteza se određuje izračunavanjem F-statistike:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)},$$

gde je $TSS = \sum(y_i - \bar{y})^2$, a $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Pokazuje se da kada ne postoji veza između odgovora i prediktora, moglo bi se očekivati da F-statistika ima vrednost blizu 1. Sa druge strane, ako je Ha tačna hipoteza, očekujemo da F vrednost bude veća od 1.

F-statistika za model višestruke linearne regresije dobijena regresiranjem prodaje na radio, TV i novine prikazana je ranije. U ovom primeru F-statistika je 570 (ceo model). Budući da je ovo daleko veće od 1, ova vrednost pruža uverljive dokaze protiv nulte hipoteze H_0. Drugim rečima, velika F-statistika sugeriše da bar jedan od reklamnih medija mora biti povezan sa prodajom. Za bilo koju datu vrednost n (**broj observacija**) i p (**broj prediktora**), bilo koji statistički softverski paket se može koristiti za izračunavanje p-vrednosti povezane sa F-statistikom pomoću ove raspodele. Na osnovu ove p-vrednosti možemo utvrditi da li ćemo odbiti nultu hipotezu ili ne.

U (5) testiramo H_0 da su svi koeficijenti nula. Međutim, ponekad želimo da testiramo da li je određeni podskup q koeficijenata nula:

$$H_0: \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0 \qquad (6)$$

U ovom slučaju fitujemo drugi model koji koristi sve promenljive osim onih poslednjih q.

opciono

Primetite da je u *MLR* tabeli 1 za svaki pojedinačni prediktor data t-statistika i prikazana je p-vrednost. Oni pružaju informacije o tome da li je svaki pojedinačni prediktor povezan sa odgovorom, nakon prilagođavanja za ostale prediktore. Pokazuje se da je svaki od t-testova tačno ekvivalentan F-testu koji izostavlja tu pojedinačnu promenljivu iz modela, a sve ostale ostavlja u modelu - tj. $q = 2$ u (6). Dakle, on izveštava o delimičnom efektu dodavanja te promenljive modelu.

S obzirom na ove pojedinačne vrednosti p za svaku promenljivu, zašto treba da gledamo ukupnu F-statistiku?

Na primer, posmatrajmo primer u kome je $p = 100$ (broj prediktora) i $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ je tačno, tako da nijedna promenljiva nije istinski povezana sa odgovorom. U ovoj situaciji, oko 5% p-vrednosti koje su povezane sa svakom promenljivom imaće vrednost ispod 0.05 (5%). Drugim rečima, očekujemo da ćemo videti približno pet malih p-vrednosti čak i u odsustvu bilo kakve prave povezanosti između prediktora i odgovora.

Međutim, F-statistika nema taj problem, jer se prilagođava broju prediktora. Dakle, ako je H_0 tačno, postoji samo 5% šanse da F-statistika rezultira p-vrednošću ispod 0.05, bez obzira na broj prediktora ili broj posmatranja.

Korišćenje F-statistike za testiranje bilo kakve povezanosti između prediktora i odgovora je moguć kada je broj prediktora p relativno mali, a sigurno mali u poređenju sa n. Međutim, ponekad imamo vrlo veliki broj promenljivih. Ako je $p > n$, tada imamo više koeficijenata β_j za procenu nego observacija iz kojih se one mogu proceniti. U ovom slučaju ne možemo koristiti model višestruke linearne regresije koristeći metodu najmanjih kvadrata, tako da se i F-statistika ne može koristiti, kao ni većina ostalih koncepata koje smo do sada opisali.

2. Određivanje bitnih prediktora (eng. *Feature Selection*)

Prvi korak u višestrukoj regresionoj analizi je izračunavanje F-statistike i određivanje pridružene p-vrednosti. Ako na osnovu te p-vrednosti možemo da zaključimo da je bar jedan od prediktora povezan sa odgovorom, onda je prirodno zapitati se koji su to prediktori!

Mogli bismo da pogledamo pojedinačne vrednosti p kao u *Tabeli 1*, ali kao što smo već objasnili, ako je p veliko, verovatno ćemo napraviti i neke loše prepostavke.

Moguće je da su svi prediktori povezani sa odgovorom, ali češći je slučaj da se odgovor odnosi samo na podskup prediktora. Zadatak utvrđivanja koji su prediktori povezani sa odgovorom, kako bi se uklopili u jedan model koji uključuje samo te prediktore, naziva se odabirom promenljivih (eng. *feature selection*).

U idealnom slučaju, želeli bismo da izvršimo izbor promenljivih isprobavanjem velikog broja različitih modela, od kojih svaki sadrži drugačiji podskup prediktora. Ako je $p = 30$, tada je broj modela $2^p = 2^{30} = 1.073.741.842$ modela.

Mogu se koristiti razne statistike za procenu kvaliteta modela. Tu spadaju *Mallow's Cp*, *Akaike informativni kriterijum (AIC), Bajesov informativni kriterijum (BIC)* i *prilagođeni* R^2(adjusted R^2).

Postoje tri klasična pristupa za ovaj zadatak:

- *Forward selection*
- *Backward selection*
- *Mixed selection* – kombinacija prva dva pristupa.

3. Tačnost modela

Dve od najčešćih numeričkih mera za merenje tačnosti modela su *RSE* i R^2 - objašnjava udeo objašnjene varijanse. Ove veličine se izračunavaju i tumače na isti način kao i za prostu linearnu regresiju.

Može se pokazati da će se R^2 uvek povećavati kada se modelu doda više promenljivih, čak iako su te promenljive samo slabo povezane sa odgovorom.

Činjenica da dodavanje novinskog oglašavanja modelu koji sadrži samo TV i radio oglašavanje dovodi do samo malog povećanja R^2 pruža dodatne dokaze da se novine mogu izbaciti iz modela. U osnovi, novine ne pružaju stvarno poboljšanje u modelu koji odgovara uzorcima za obuku, a njegovo uključivanje će verovatno dovesti do loših rezultata na nezavisnim test uzorcima zbog *overfitting*-a.

Nekada se može desiti da se *RSE* vrednost poveća kada se doda novi prediktor. Kako se *RSE* može povećati kada se novi prediktori dodaju modelu s obzirom da se *RSS* mora smanjiti? Generalno, *RSE* se definiše kao:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

Dakle, modeli sa više promenljivih mogu imati veći *RSE* ako je smanjenje *RSS*-a malo u odnosu na povećanje p.

Na nekom od narednih predavanja biće reči o proširenju linearnog modela kako bi se mogli prilagoditi **sinergijski efekti korišćenjem interakcije** među njima (eng. synergy or interaction effect).

4. Predikcija

Kada je model na osnovu linearne regresije kreiran, sada veoma lako možemo predvideti izlaz Y na osnovu vrednosti prediktora $X_1, X_2, \dots, X_p$. Međutim, postoje tri vrste neodređenosti koje su povezane sa ovom predikcijom.

1. Određeni koeficijenti $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ su aproksimacije za prave koeficijente $\beta_0, \beta_1, \beta_2, \dots, \beta_p$. Drugim rečima, *ravan najmanjih kvadrata*

   $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

   je samo aproksimacija prave regresione ravni

   $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Netačnost u proceni koeficijenta povezana je sa *reducibilnom greškom*. Sada se može izračunati **interval poverenja** (eng. *confidence interval*) kako bi se odredilo koliko će $\hat{Y}$ biti blizu $f(X)$.

2. U praksi linearan model za $f(X)$ je aproksimacija realnog problema, tako da postoji dodatni izvor za reducibilnu grešku koji se naziva *model bias*. Dakle, kada koristimo linearni model, mi u stvari procenjujemo najbolju linearnu aproksimaciju za pravu površ.
3. Štaviše, i da znamo prave vrednosti koeficijenata $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ output se ne može predvideti pefrektno zbog greške e. Ovo se naziva *ireducibilnom greškom*. Koliko se $\hat{Y}$ razlikuje od Y? **Intervali predikcije** (eng. *prediction intervals*) se koriste da bi se dao odgovor na ovo pitanje. Intervali predikcije su širi od intervala poverenja, jer u obzir uzimaju obe vrste greške. *Formule za intervale poverenja i predikciju nećemo navoditi, jer se lako mogu pronaći na webu (opciono)*.

Interval poverenja koristimo za kvantifikovanje neodređenosti koja se odnosi na prosečnu prodaju u velikom broju gradova. Na primer, s obzirom na to da se potroši \$100 000 na TV oglašavanje, a \$20 000 na radio oglašavanje u svakom gradu, interval poverenja od 95% iznosi [10.985, 11.528]. To tumačimo tako da će 95% interval ovog oblika da sadrži pravu vrednost za $f(X)$.

S druge strane, interval predikcije se može se koristiti za kvantifikaciju neodređenosti predviđanja koje se odnosi na prodaju za određeni grad. S obzirom na to da se \$100 000 potroši na TV oglašavanje, a \$20 000 na radio oglašavanje u tom gradu, interval predviđanja od 95% iznosi [7.930, 14.580]. To tumačimo tako da će 95% interval ovog oblika da sadrži pravu vrednost Y za ovaj grad.

% R code

```
model_tv_radio = lm(sales~TV+radio, data=advertising)
```

```
a) predict(tv_radio, advertising[c(1:5), c("TV", "radio")], interval = "confidence")
b) predict(tv_radio, advertising[c(1:5), c("TV", "radio")], interval = " prediction")
```