Az NFL irányítóinak teljesítménye

Heller Farkas Szakkollégium

Ökonometria I. kurzus

Juhász Kristóf

FM9AGF

2016-17 őszi félév

Tartalomjegyzék

1. Az adatbázis bemutatása

A megfelelő adatbázis megtalálásával viszonylag sok időt töltöttem, ugyanis szakítani akartam a közgazdasági témákkal, de mindenképpen olyan területet szerettem volna vizsgálni, ami számomra érdekes és újdonságot is rejt egyben. Így végül a sport témán belül az NFL mellett döntöttem.

Az adatbázis az elmúlt tíz év NFL-ben játszó/játszott irányítóinak (quarterback) meccsenkénti teljesítményét tartalmazza. Mivel nekem sok változó esetében az összehasonlíthatóság miatt átlagos teljesítményekre volt szükségem, így ezeket a változtatásokat az adatbázis irreleváns változóitól való megtisztítása után elvégeztem. Az így kapott 8 változó, amely egy-egy irányítót jellemez:

Változó	**Leírás**	**Átalakítás**
points	Szerzett pontszám	átlagos
cpera	Összes célba-ért passza / Összes passzkísérlete	kalkulált
yard	„Megtett” yardok száma	átlagos
int	Ellenfél által elhalászott/elkapott passzainak száma	átlagos
long	Karrierjének leghosszabb dobása	max
sack	Leszereléseinek száma miközben a labda a kezében volt	átlagos
loss	Vesztett „down”-ok száma (ilyenkor az ellenfél következik támadni)	átlagos
home	Hazai meccsei / Összes meccse	kalkulált

**az átlagolt értékeknél a súly az irányító által játszott meccsek száma, azaz az értékek egy meccs alatti átlagos teljesítmények*

Ezen kívül még megjegyezném, hogy az adatbázist azoktól a soroktól (irányítóktól) is megtisztítottam, akik átlagosan egy meccs alatt 100 yardnál kevesebbet tettek meg, hiszen ezek nagy többsége irányító posztra becserélt rugó játékos vagy sérülés miatti csere. Ez egy nagyon kicsit torzíthatja az eredményeket, de úgy gondoltam, hogy „sikeres” és a modellem szerint relevánsnak számító irányítónak csak az mondható, aki meccsenként átlagosan legalább 100 yardot képes megtenni.

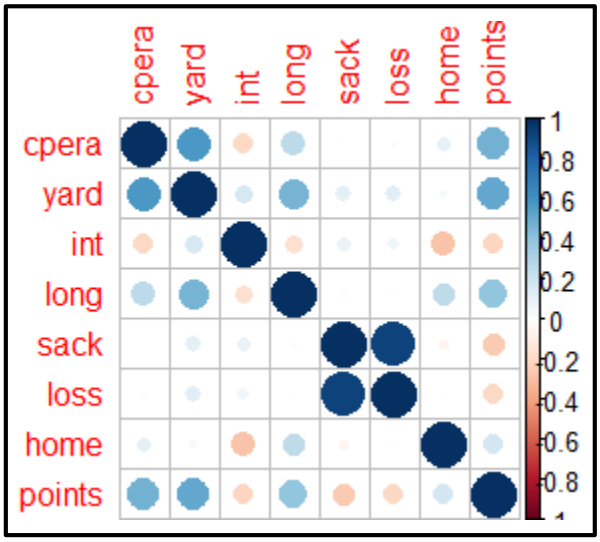
2. Modelldiagnosztika

Bennem legérdekesebb vizsgálható kérdésként az vetődött fel, hogy a meccsenként szerzett átlagos pontszámot egy-egy irányító esetében hogyan befolyásolja a többi fent felsorolt változó (magyarázóváltozók), így a points változót választottam eredményváltozónak. Az R-ben lefuttatott első summary eredménye a következő lett:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.30164    3.76014   1.942 0.053415 .
cpera       13.95502    6.45939   2.160 0.031806 *
home         1.95429    2.12333   0.920 0.358364
long         0.02420    0.01673   1.447 0.149293
int         -2.65580    0.72395  -3.669 0.000305 ***
yard         0.04282    0.00683   6.269 1.86e-09 ***
sack        -2.14753    0.81283  -2.642 0.008825 **
loss         0.06626    0.11957   0.554 0.580043
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.093 on 223 degrees of freedom
Multiple R-squared:  0.4674,    Adjusted R-squared:  0.4506
F-statistic: 27.95 on 7 and 223 DF,  p-value: < 2.2e-16
```

Az így kapott modell magyarázóereje nem túl jó, és pár magyarázóváltozó relevanciája is megkérdőjelezhető, ezért a korrelációs mátrixot kezdtem el vizsgálni, mert gyanítottam, hogy egy-két változónál a korreláció lehet az alapvető probléma:



Látszik, hogy a sack és a loss változók között nagyon erős korreláció van (szám szerint majdnem 93%), így a kettő közül az egyiket biztosan ki kell zárnom. Először a loss változó elimánálásával (csak sack-el korrigált R^2=0.4523), majd a sack változó elimánálásával (csak loss-szal korrigált R^2=0.436) futtattam újra a modellt. A kapott determinációs együtthatók alapján úgy döntöttem, hogy a sack változót tartom meg a kettő közül, mert többet tesz hozzá a modell magyarázó erejéhez. (Egyébként ez az erős korreláció azért állhat fent, mert ha az irányító sack-et kap, akkor az loss of down-al jár, azaz elveszti a támadás jogát, de loss of down-t még sok más egyéb okból is kaphat az irányító. Sajnos előre nem gondoltam, hogy ennyire gyakori a sack-ért kapott loss.)

Ezután a backward elimination módszert követve, mindig a legnagyobb p értékű magyarázó változót, amely nem éri el legalább a hagyományos 95%-os konfidenciaszint környékét, sorra kizártam. Minden magyarázó változó kizárás után újrafuttattam a modellt a korrigált R^2 és a p értékek változásának nyomon-követése miatt, hogy mindig ténylegesen azt a változót zárjam ki, amelyik a legkevesebbet ront a modell magyarázóerején. Végül ezzel a fokozatos módszerrel erre a summary-re jutottam:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.302098   3.456273   2.981  0.00319 **
yard         0.048473   0.005976   8.111 3.19e-14 ***
cpera       12.270073   6.376344   1.924  0.05557 .
int         -3.189946   0.675401  -4.723 4.08e-06 ***
sack        -1.777464   0.304834  -5.831 1.89e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.101 on 226 degrees of freedom
Multiple R-squared:  0.4575,    Adjusted R-squared:  0.4479
F-statistic: 47.65 on 4 and 226 DF,  p-value: < 2.2e-16
```

Látszik, hogy a cpera változó még így is kilóg a 95%-os konfidencia intervallumból (és nem is túl releváns a modell magyarázóereje szempontjából), azonban annyira határon van (p érték és relevancia szempontjából), hogy mielőtt eliminálnám, mindenképp multkikollinearitást és homoszkedaszticitást vizsgálnék.

3. Modellspecifikáció tesztelése

A modellspecifikáció tesztelésére a Ramsay RESET test-et futtattam, amely azt vizsgálja, hogy megfelelő-e a lineáris forma vagy a változóknak egy nemlineáris kombinációja magyarázná-e jobban az eredményváltozót. Ilyenkor H_0 hipotézisünk az, hogy az összes nemlineáris tag koefficiense 0. Az eredményből látható, hogy H_0-t elég nagy biztonsággal elfogadhatjuk, azaz a lineáris forma megfelelő.

```
RESET = 0.99036, df1 = 2, df2 = 224, p-value = 0.3731
```

4. Multikollinearitás vizsgálata

Mindenekelőtt egy korrelációs vizsgálattal kezdeném, de ezt már a fentiekben megtettem, és segítségével egy változót sikeresen ki is tudtam zárni.

A multikollinearitás vizsgálatához a VIF mutatót hívtam segítségül. Ennek nagysága egyes magyarázóváltozókra megmutatja, hogy a tényleges varianciájukat hányszorosára nagyítja fel a többi magyarázóváltozóval való együttmozgás hatása. A VIF mutatók a következők lettek:

yard	cpera	int	sack
1.732118	1.739704	1.207125	1.022392

Mivel alapvetően minden magyarázóváltozó VIF mutatója 2 alatt (és természetesen 1 fölött) van, azt mondhatjuk, hogy multikollinearitás nem áll fenn.

5. Heteroszkedaszticitás vizsgálata

A heteroszkedaszticitás vizsgálatát a Breusch-Pagan próbával végeztem el. Null-hipotézisünk itt azt jelenti, hogy a hibatagok szórása állandó, azaz homoszkedaszticitás áll fenn.

```
BP = 30.94, df = 4, p-value = 3.149e-06
```

A próba szerint azonban H0 majdnem teljesen biztonsággal elvethető, így a modellben heteroszkedaszticitás van, amit kezelnünk kell. Így hát R-rel megbecsültem a heteroszkedaszticitásra konzisztens kovariancia-mátrixot. A főátlójában lévő varianciák gyökeiként kaphatjuk meg a további teszteléshez szükséges standard hibákat. A kapott standard hibákkal ezután t-próbákat csináltam, hogy feltárjam mégis mekkora problémát jelent a változók magyarázó erejében a heteroszkedaszticitás fennállása.

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 10.3020980  5.4129651  1.9032   0.05828 .
yard         0.0484730  0.0099716  4.8611 2.185e-06 ***
cpera       12.2700729 10.2716734  1.1946   0.23351
int         -3.1899460  1.4518935 -2.1971   0.02903 *
sack        -1.7774636  0.3303079 -5.3812 1.846e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ebből látszik, hogy a kezdeti modelldiagnosztikában kapott summary szerint is már elvetendőnek minősített cpera változót mindenképp eliminálni kell. E változó nélkül újra elvégeztem a t próbákat, amelynek a végeredménye mutatja, hogy már nem áll fent jelentős heteroszkedaszticitás:

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 16.6453206  1.3422695 12.4009 < 2.2e-16 ***
yard         0.0557359  0.0053472 10.4234 < 2.2e-16 ***
int         -3.6793969  1.2178405 -3.0212  0.002806 **
sack        -1.8157369  0.3366545 -5.3935 1.732e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. A felépített modell

A megmaradt magyarázóváltozókra újrafuttattam egy summary-t, egy RESET tesztet és egy VIF tesztet a multikollinearitás vizsgálatára.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.645321   1.045320  15.924  < 2e-16 ***
yard         0.055736   0.004661  11.957  < 2e-16 ***
int         -3.679397   0.629391  -5.846 1.74e-08 ***
sack        -1.815737   0.305990  -5.934 1.10e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.119 on 227 degrees of freedom
Multiple R-squared:  0.4486,    Adjusted R-squared:  0.4413
F-statistic: 61.56 on 3 and 227 DF,  p-value: < 2.2e-16
```

```
data:  fit
RESET = 1.3927, df1 = 2, df2 = 225, p-value = 0.2505
```

```
    yard      int     sack
1.041316 1.035928 1.018040
```

A meggyőző eredmény után a biztonság kedvéért egy ANOVA táblát is kiírattattam az R-rel. A kapott F értékek már ránézésre is egytől egyig elég magasnak mondhatóak (ekkora elemszám mellett), így ez is igazolja a változók relevanciáját.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
yard          1 1071.0  1071.0  110.08  < 2e-16 ***
int           1  383.3   383.3   39.39 1.74e-09 ***
sack          1  342.6   342.6   35.21 1.10e-08 ***
Residuals   227 2208.5     9.7
```

Az elvégzett vizsgálatok alapján tehát kijelenthető, hogy az átlagosan megtett yardok száma (yard), az átlagosan elhalászott indítások száma (int) és a labdával a kézben való leszerelések száma (sack) 44,13%-ban magyarázza a meccsenként átlagosan szerzett pontszámot (points). Sajnos a modellnek csak gyenge közepes magyarázóereje van, mivel nem tartalmaz minden releváns faktort a szerzett pontok becsléséhez. Így ténylegesen, a való életben aligha lehetne használni a modellt a pontszám becslésére csupán ennyi magyarázóváltozó felhasználásával, azonban most mindössze ennyi állt rendelkezésemre.

7. A regressziós egyenlet

7.1. A regressziós egyenlet felírása

A fent említett summary funkcióval kapott lekérdezés tartalmazza a koefficienseket, valamint a RESET teszt alapján megfelelő a lineáris forma, így már könnyedén felírható a regressziós egyenlet:

$$points = 16{,}645321 + 0{,}055736 * yard - 3{,}679397 * int - 1{,}815737 * sack$$

7.2. A regressziós egyenlet paramétereinek értelmezése

A konstans tag kissé nehezen értelmezhető, hiszen ez azt jelenti, hogy ha minden egyéb változó értéke nulla, akkor a megszerzett pontok összege várhatóan 16-17 lesz. Az azonban, hogy egy irányítót egyáltalán ne szereljenek, egy passzát se kapják el, sőt az irányító egy yardot se haladjon, és így összegyűjtsön 16-17 pontot szinte kizárt (elméletileg lehetséges lehet csak pontrúgásból pontokat gyűjteni, de nem sokáig lenne az a játékos irányító, aki nem tud egy yardot se nyerni a csapatának).

A yard változó együtthatója azt mutatja meg, hogy ha egy irányító 10 yardot (10 yardonként szokás mérni) nyer csapatának, az, ceteris paribus, kicsit több mint fél ponttal növeli előreláthatólag az irányító által átlagosan szerzett pontok számát.

Az int és sack változók értelmezése is hasonlóképpen történik.

8. Útelemezés

Az útelemzésnél az átlagosan egy meccs alatt megtett yardok számának teljes hatását bontottam fel közvetlen és közvetett tényezőkre.

A yard változó teljes hatását úgy kapom meg, ha egy olyan modellt tesztelek, amiben csupán ezzel az egy változóval magyarázom az eredményváltozót (points).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.153686   0.946539  11.784   <2e-16 ***
yard         0.047928   0.005242   9.142   <2e-16 ***
```

Így, ha csak a megtett yardokkal magyaráznám az elért pontot, azt kapnám, hogy 10 megtett yard várhatóan 0,48-al növeli az irányító által szerzett pontok számát egy meccsen.
A közvetlen hatás az eredeti modellből vett koefficiense a yard magyarázóváltozónak (= 0,055736), amelynek jelentését fentebb már kifejtettem.
A közvetett hatást 3 magyarázó változónál már viszonylag bonyolultabb számolni, azonban a teljes hatásból a közvetlen hatás értékét kivonva megkaphatjuk könnyedén az egy meccs alatt átlagosan megtett yardok közvetett hatását, ami -0,00781.

9. Előrejelzés

Mindenekelőtt megemlíteném még egyszer, hogy sajnos a modellnek a magyarázóereje csupán közepes, így könnyen megeshet, hogy a becslések korántsem fedik a tényleges valós kimeneteleket (adott paraméterek mellett).

9.1. Pontbecslés

Pontbecslést úgy végezhetünk, ha a már korábban említett regressziós egyenletben a magyarázóváltozóknak egy-egy fix értéket adunk, egy-egy értéket helyettesítünk be, majd megvizsgáljuk, milyen értéket vesz fel a behelyettesített értékek mellett az eredményváltozó. Hogy ne rugaszkodjunk el túlságosan a valóságtól, legyen az átlagosan megtett yardok száma (yard) 250, az ellenfél által átlagosan elcsent passzok száma (int) 1 és az átlagosan kapott (labdával a kézben lévő) szerelések száma (sack) 3. Így az eredményváltozónk, azaz az átlagosan szerzett pontok száma, a következőképpen alakul:

$$16{,}645321 + (0{,}055736 * 250) - (3{,}679397 * 1) - (1{,}815737 * 3) =$$

```
> predict(fit, data.frame(yard=250,int=1,sack=3)) =

21.45269
```

Azaz a fent említett rögzített paraméterek mellett várhatóan az irányító által meccsenként átlagosan szerzett pontszám 21,5 körül lesz.

9.2. Intervallumbecslés

Az intervallumbecslést a már megadott paraméterekkel a pontbecsléshez hasonlóan számoltam.

```
> predict(fit, data.frame(yard=250,int=1,sack=3),interval="predict")
       fit      lwr      upr
1 21.45269 15.23566 27.66972
```

Ez azt jelenti, hogy a rögzített paraméterek mellett, 95%-os valószínűséggel az irányító által meccsenként átlagosan szerzett pontszám kb. 15 és 28 közé fog esni. A becslési intervallum a modell közepes magyarázóereje miatt ilyen tág.

10. Összegzés

Az adatbázis választásom nem volt a legmegfelelőbb, de mivel rendhagyó adatokat szerettem volna elemezni, így vállaltam a velük járó kockázatot. A tanult módszereket felhasználva fokozatosan tisztítottam az adataimat, míg a végső modellemhez nem érkeztem, majd a regressziós egyenlet felírása és a változók elemzése után elvégeztem az előrejelzést. Ugyan a modellemet nem fogja megvásárolni a Szerencsejáték Zrt. a Tippmix odds-ok kalkulálásához, de az órán elsajátított módszereket a gyakorlatba ültetve tudásomat sikeresen bővítettem az ökonometria tárgyterületén (és az amerikai focién is).