Tjedan 13: Linearna regresija

Predviđanje i objašnjavanje s modelima

2025-05-24

Table of contents

```
library(tidyverse)
```

Ishodi učenja

Nakon ovog predavanja moći ćete:

1. Objasniti razliku između korelacije i regresije.
2. Provesti jednostavnu linearnu regresiju u R-u i interpretirati koeficijente.
3. Interpretirati R-kvadrat kao mjeru kvalitete modela.
4. Provjeriti pretpostavke linearne regresije dijagnostičkim grafovima.
5. Provesti višestruku regresiju s više prediktora i interpretirati parcijalne koeficijente.
6. Usporediti modele pomoću R-kvadrata, prilagođenog R-kvadrata i AIC-a.
7. Prepoznati uobičajene probleme (multikolinearnost, utjecajne točke, nelinearnost).
8. Napisati kompletni izvještaj regresijske analize.

1 Što pokreće angažman?

Zamislite sljedeću situaciju. Radite kao analitičarka društvenih mreža za srednje veliku medijsku kuću. Vaša šefica dolazi s pitanjem koje zvuči jednostavno poput "Koji faktori utječu na angažman naših Instagram objava?" Želi znati je li stvar u duljini teksta, broju hashtagova, tipu sadržaja, pozivu na akciju, ili u nečem sasvim drugom. I još važnije, želi konkretne preporuke — što da radimo više, a što manje?

Do sada ste u kolegiju naučili uspoređivati grupe. Hi-kvadrat test govori vam postoji li veza između kategoričkih varijabli. T-test uspoređuje prosjeke dviju grupa. ANOVA uspoređuje više grupa odjednom. Ali nijedno od toga ne odgovara na pitanje vaše šefice. Ona ne pita "razlikuju li se grupe." Umjesto toga, pita koliko svaki faktor doprinosi angažmanu, u kojem smjeru, i koliko dobro možemo predvidjeti angažman na temelju tih faktora.

Za to nam treba regresija. Regresija je — u najjednostavnijem smislu — alat koji modelira odnos između jedne ili više nezavisnih varijabli (koje zovemo prediktorima) i jedne zavisne varijable (koju zovemo ishodom). Umjesto da samo kaže "postoji razlika", regresija kvantificira — za svaki dodatni hashtag, angažman se mijenja za toliko i toliko. To je razlika između "hashtagovi su važni" i "svaki dodatni hashtag smanjuje angažman za 0.15 postotnih bodova, kontrolirajući za ostale faktore."

Radimo s datasetom od 500 Instagram objava jednog poslovnog profila. Svaka objava ima zabilježen engagement rate (postotak pratitelja koji su reagirali), duljinu teksta, broj hashtagova, broj oznaka drugih profila, tip sadržaja (slika, video, carousel, reel), temu i informaciju o tome je li uključen poziv na akciju (CTA).

```r
posts <- read_csv("../resources/datasets/social_engagement.csv")
glimpse(posts)
```

```
Rows: 400
Columns: 14
$ post_id         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
$ day             <chr> "utorak", "ponedjeljak", "petak", "nedjelja", "subota"~
$ time_slot       <chr> "18-21", "09-12", "21-24", "12-15", "15-18", "18-
21", ~
$ content_type    <chr> "foto", "tekst", "carousel", "foto", "carousel", "reel~
$ topic           <chr> "iza_kulisa", "proizvod", "zabava", "zabava", "proizvo~
$ text_length     <dbl> 290, 34, 35, 162, 240, 189, 300, 228, 242, 97, 136, 17~
$ num_hashtags    <dbl> 0, 17, 19, 14, 20, 9, 25, 21, 6, 12, 0, 23, 27, 27, 20~
$ has_cta         <dbl> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, ~
$ num_mentions    <dbl> 3, 1, 1, 4, 0, 4, 4, 1, 1, 1, 2, 4, 1, 4, 4, 4, 3, 0, ~
$ followers       <dbl> 15095, 15312, 14749, 14728, 15336, 14722, 15168, 14805~
$ engagement_rate <dbl> 4.93, 2.26, 7.32, 5.58, 4.03, 6.58, 3.51, 6.74, 2.03, ~
$ likes           <dbl> 594, 284, 860, 583, 500, 692, 447, 816, 212, 548, 476,~
$ comments        <dbl> 120, 76, 174, 81, 93, 138, 88, 195, 45, 150, 87, 33, 4~
$ shares          <dbl> 78, 32, 115, 91, 70, 124, 41, 59, 45, 92, 80, 34, 25, ~
```

2 Od korelacije do regresije

Krenimo od poznatog terena. Korelaciju već znate — ona mjeri jačinu i smjer linearne veze između dviju varijabli. Pearsonov r kreće se od -1 (savršena negativna veza) preko 0 (nema linearne veze) do +1 (savršena pozitivna veza).

Regresija ide korak dalje. Dok korelacija samo kaže "ove dvije varijable su povezane", regresija definira jednadžbu pravca koja opisuje tu vezu. Ta jednadžba vam omogućuje nešto što korelacija ne može — predviđanje. Ako znate koliko hashtagova ima objava, regresija vam daje konkretnu procjenu koliki će biti njezin angažman.

Prije nego što uđemo u regresiju, pogledajmo korelacije između naših numeričkih varijabli.

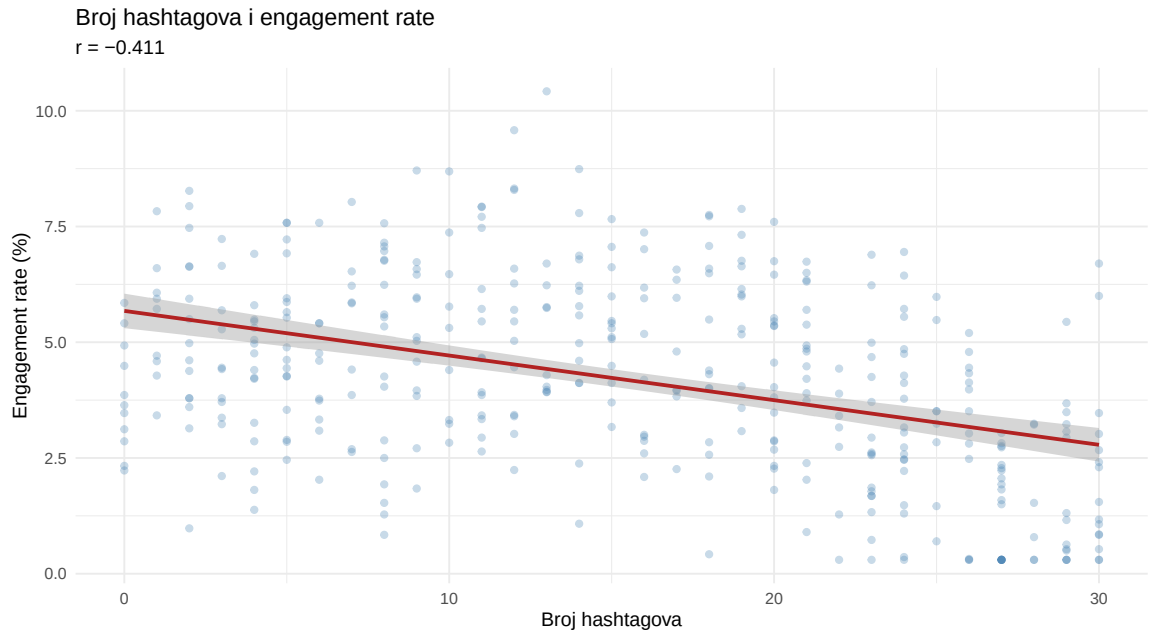
```r
# Korelacije numeričkih prediktora s engagement_rate
posts |>
  select(engagement_rate, text_length, num_hashtags, num_mentions, followers) |>
  cor() |>
  round(3)
```

```
               engagement_rate text_length num_hashtags num_mentions followers
engagement_rate           1.000       0.040       -0.411       -0.016    -
0.132
```

```
text_length          0.040       1.000       0.008       0.061    -
0.025
num_hashtags        -0.411       0.008       1.000       0.031    0.062
num_mentions        -0.016       0.061       0.031       1.000    0.092
followers           -0.132      -0.025       0.062       0.092    1.000
```

Pogledajte stupac `engagement_rate`. Broj hashtagova ima negativnu korelaciju s angažmanom (r -0.41), što znači da objave s više hashtagova u prosjeku imaju niži engagement rate. Vizualizirajmo tu vezu.

```
posts |>
  ggplot(aes(x = num_hashtags, y = engagement_rate)) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "firebrick") +
  labs(
    title = "Broj hashtagova i engagement rate",
    subtitle = paste0("r = ", round(cor(posts$num_hashtags, posts$engagement_rate), 3)),
    x = "Broj hashtagova",
    y = "Engagement rate (%)"
  ) +
  theme_minimal()
```



Crvena linija je regresijski pravac — ona predstavlja "najbolju" ravnu liniju koja prolazi kroz oblak točaka. Sivi pojas oko nje pokazuje nesigurnost te procjene (95% interval pouzdanosti za pravac).

Jedna stvar vas možda brine. Negativna korelacija sugerira da više hashtagova znači niži angažman. Ali budite oprezni s takvim zaključcima. Možda veza uopće nije linearna —

možda postoji optimalan broj hashtagova, a i premalo i previše je loše. Možda profili s više hashtagova imaju i druge karakteristike koje snižavaju angažman. To su pitanja koja ćemo istražiti kasnije u ovom predavanju.

3 Jednostavna linearna regresija

Počnimo s najjednostavnijim mogućim modelom — jednim prediktorom i jednim ishodom. To je jednostavna linearna regresija.

Ideja je intuitivna. Vi imate oblak točaka na scatterplotu i želite provući ravnu liniju kroz taj oblak tako da ona što bolje opisuje opći trend. "Što bolje" u praksi znači da je ukupna udaljenost svih točaka od linije što manja.

Matematički, ta linija izgleda ovako.

$$Y = b_0 + b_1 X + \varepsilon$$

Raspakirajmo ovo simbol po simbol. Y je vaša zavisna varijabla, ono što želite predvidjeti (u našem slučaju engagement rate). X je prediktor (recimo, duljina teksta). b_0 je odsječak, koji vam kaže koliki bi bio predviđeni engagement rate kad bi duljina teksta bila nula. b_1 je nagib, ključni broj — on govori za koliko se engagement rate mijenja kad duljina teksta poraste za jednu jedinicu. Konačno, ε je greška, rezidual, ono što model ne uspijeva objasniti. Svaka objava ima svoju priču koja nije samo u duljini teksta.

Pokrenimo regresiju u R-u. Funkcija `lm()` (linear model) traži formulu i podatke. Formula `engagement_rate ~ text_length` znači "predvidi engagement rate na temelju duljine teksta".

```r
# Jednostavna regresija: text_length -> engagement_rate
model1 <- lm(engagement_rate ~ text_length, data = posts)
summary(model1)
```

```
Call:
lm(formula = engagement_rate ~ text_length, data = posts)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9713 -1.4744 -0.0468  1.5489  6.1332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.034491   0.238994  16.881   <2e-16 ***
text_length 0.001034   0.001299   0.796    0.426
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.102 on 398 degrees of freedom
Multiple R-squared:  0.001591,  Adjusted R-squared:  -0.0009177
F-statistic: 0.6342 on 1 and 398 DF,  p-value: 0.4263
```

3.1 Kako čitati ovaj output

Output funkcije `summary()` na regresijskom modelu sadrži puno informacija, i čest je osjećaj studenta da je "sve puno zvjezdica i brojeva" te ne znaju gdje početi. Počnimo od najvažnijeg i idimo redom.

```
koef <- coef(model1)

cat("Jednadžba: engagement_rate = ", round(koef[1], 3), " + ",
    round(koef[2], 5), " * text_length\n\n", sep = "")
```

```
Jednadžba: engagement_rate = 4.034 + 0.00103 * text_length
```

```
cat("Interpretacija:\n")
```

```
Interpretacija:
```

```
cat("  Intercept (b0 = ", round(koef[1], 2), "): Ocekivani engagement rate\n", sep = "")
```

```
  Intercept (b0 = 4.03): Ocekivani engagement rate
```

```
cat("  kad je text_length = 0 (teorijska vrijednost, nema prakticnog znacenja).\n\n")
```

```
  kad je text_length = 0 (teorijska vrijednost, nema prakticnog znacenja).
```

```
cat("  Slope (b1 = ", round(koef[2], 4), "): Za svaki dodatni znak u tekstu,\n", sep = "")
```

```
  Slope (b1 = 0.001): Za svaki dodatni znak u tekstu,
```

```
cat("  engagement rate se mijenja za ", round(koef[2], 4), " postotnih bodova.\n\n", sep =
```

```
  engagement rate se mijenja za 0.001 postotnih bodova.
```

```r
# R-kvadrat
r2 <- summary(model1)$r.squared
cat("R-kvadrat:", round(r2, 4), "\n")
```

```
R-kvadrat: 0.0016
```

```r
cat("Interpretacija:", round(r2 * 100, 1), "% varijabilnosti u engagement rateu\n")
```

```
Interpretacija: 0.2 % varijabilnosti u engagement rateu
```

```r
cat("je objasnjeno duljinom teksta. To je vrlo malo.\n")
```

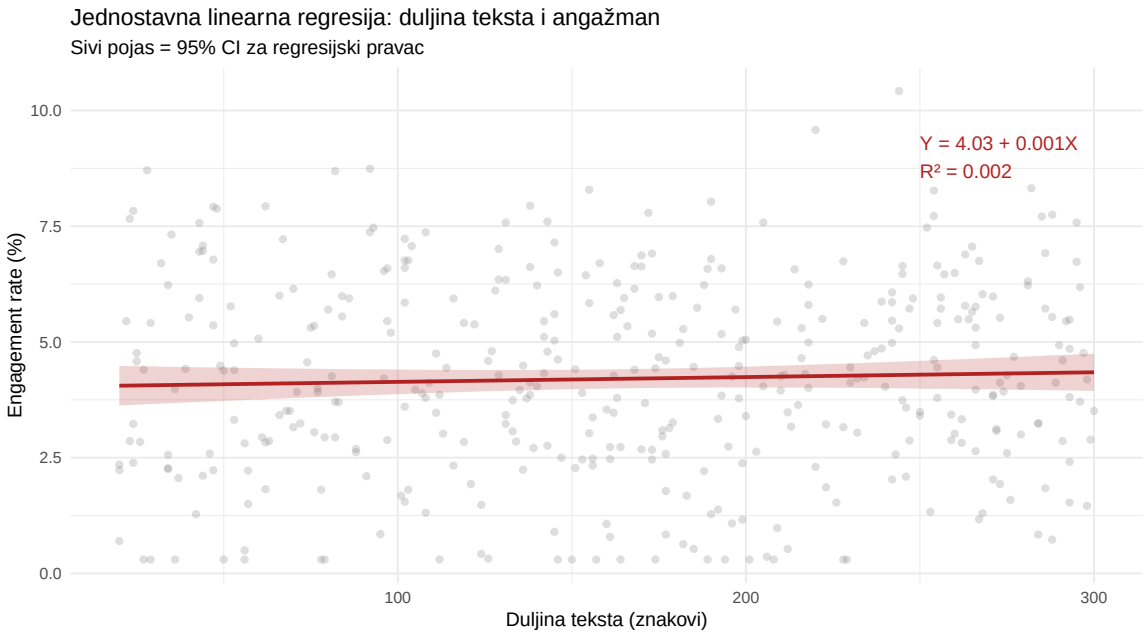
```
je objasnjeno duljinom teksta. To je vrlo malo.
```

Koeficijenti su dva broja koja definiraju vašu liniju. Odsječak (intercept, b_0) vam kaže predviđeni engagement rate kad je duljina teksta nula. U praksi, nitko ne objavljuje objavu bez teksta, pa taj broj nema praktičnu interpretaciju, ali je potreban da definira liniju. Nagib (slope, b_1) je ono što vas zapravo zanima — za svaki dodatni znak u tekstu, engagement rate se mijenja za toliko postotnih bodova.

R-kvadrat odgovara na jedno važno pitanje — koliki udio ukupne varijabilnosti u engagement rateu objašnjava naš model? Vrijednost je niska. Duljina teksta sama jednostavno nije dobar prediktor angažmana. To ima smisla jer o angažmanu odlučuje puno više faktora od duljine teksta. Trebat će nam više prediktora.

3.2 Vizualizacija regresijskog pravca

```r
posts |>
  ggplot(aes(x = text_length, y = engagement_rate)) +
  geom_point(alpha = 0.25, color = "grey50") +
  geom_smooth(method = "lm", se = TRUE, color = "firebrick", fill = "firebrick", alpha = 0
  annotate("text", x = 250, y = 9,
           label = paste0("Y = ", round(koef[1], 2), " + ", round(koef[2], 4), "X\nR² = ",
           color = "firebrick", hjust = 0) +
  labs(
    title = "Jednostavna linearna regresija: duljina teksta i angažman",
    subtitle = "Sivi pojas = 95% CI za regresijski pravac",
    x = "Duljina teksta (znakovi)",
    y = "Engagement rate (%)"
  ) +
  theme_minimal()
```



Pogledajte koliko je oblak točaka razbacanih daleko od linije. To je vizualna manifestacija niskog R-kvadrata — linija postoji, ali objašnjava samo mali dio priče. Većina varijabilnosti dolazi od faktora koje ovaj model ne uključuje.

4 Što su reziduali?

Svaka točka na grafu ima svoju predviđenu vrijednost (točku na liniji) i svoju stvarnu vrijednost (točku u oblaku). Razlika između te dvije vrijednosti zove se rezidual.

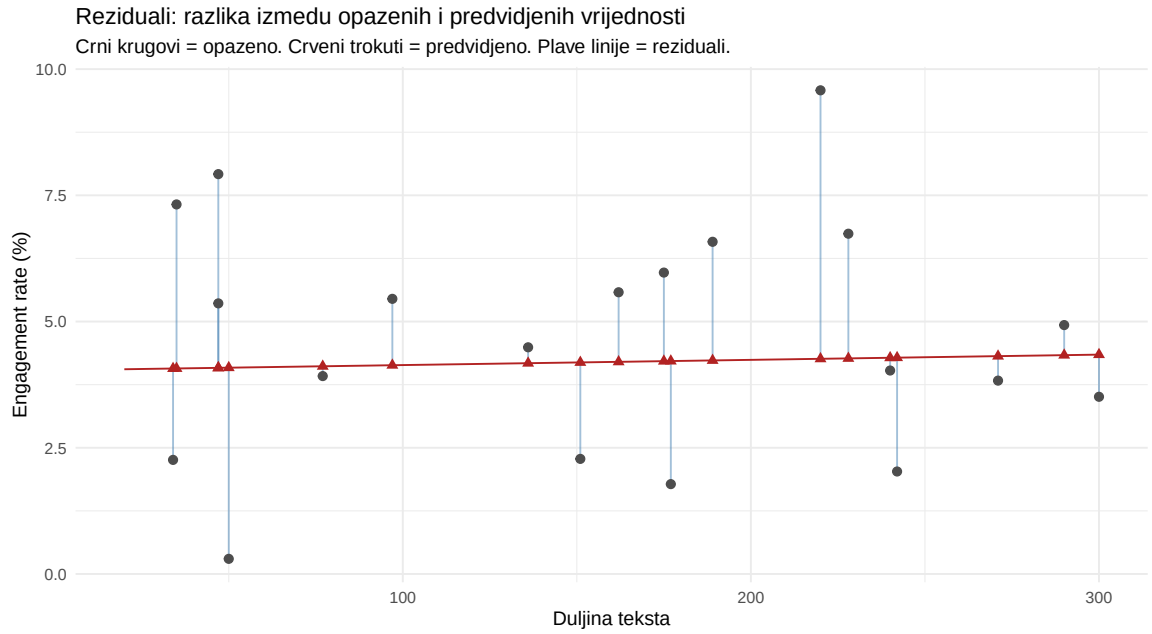
$$e_i = Y_i - \hat{Y}_i$$

U prijevodu, rezidual za i-tu objavu jednak je njezinoj stvarnoj engagement stopi minus onome što je model predvidio. Ako je rezidual pozitivan, objava je imala bolji angažman nego što je model očekivao. Ako je negativan, lošiji.

Regresija traži liniju koja minimizira sumu kvadriranih reziduala. Ova metoda zove se OLS (Ordinary Least Squares, metoda najmanjih kvadrata). Zašto kvadriramo? Jer bismo inače imali pozitivne i negativne reziduale koji bi se međusobno poništavali. Kvadriranje osigurava da su sva odstupanja pozitivna, a kao bonus, veća odstupanja kažnjavaju se proporcionalno više.

```
# Dodajmo predvidjene vrijednosti i reziduale u podatke
posts_pred <- posts |>
  mutate(
    predicted = predict(model1),
    residual = residuals(model1)
```

```r
  )

# Prikaz reziduala za prvih 20 objava
posts_pred |>
  slice(1:20) |>
  ggplot(aes(x = text_length, y = engagement_rate)) +
  geom_segment(aes(xend = text_length, yend = predicted), color = "steelblue", alpha = 0.5
  geom_point(color = "grey30", size = 2) +
  geom_point(aes(y = predicted), color = "firebrick", size = 2, shape = 17) +
  geom_smooth(data = posts, method = "lm", se = FALSE, color = "firebrick", linewidth = 0.
  labs(
    title = "Reziduali: razlika izmedu opazenih i predvidjenih vrijednosti",
    subtitle = "Crni krugovi = opazeno. Crveni trokuti = predvidjeno. Plave linije = rezid
    x = "Duljina teksta",
    y = "Engagement rate (%)"
  ) +
  theme_minimal()
```



Plave vertikalne linije na ovom grafu su reziduali. Svaka linija povezuje stvarnu vrijednost neke objave (crni krug) s njezinom predviđenom vrijednošću na regresijskom pravcu (crveni trokut). Kraće linije znače bolje predviđanje. Duže linije znače da je model za tu objavu značajno pogriješio.

Reziduali nisu samo mjera pogreške. Oni su dijagnostički alat. Ako ih pažljivo proučimo, mogu nam otkriti različite probleme s modelom uključujući nelinearnost, nejednakomjernu varijabilnost, ili utjecajne točke koje iskrivljuju cijelu analizu.

5 Pretpostavke linearne regresije

Svaki statistički test ima pretpostavke, i regresija nije iznimka. Postoje četiri ključne pretpostavke koje moraju biti barem približno zadovoljene da bismo mogli vjerovati našim rezultatima.

Linearnost — veza između prediktora i ishoda mora biti linearna. Ako je stvarna veza zakrivljena, a mi joj pokušavamo prilagoditi ravnu liniju, naši koeficijenti bit će pristrani.

Nezavisnost reziduala — reziduali jednog opažanja ne smiju biti povezani s rezidualima drugog. Ovo je obično zadovoljeno ako su opažanja prikupljena neovisno jedno o drugom.

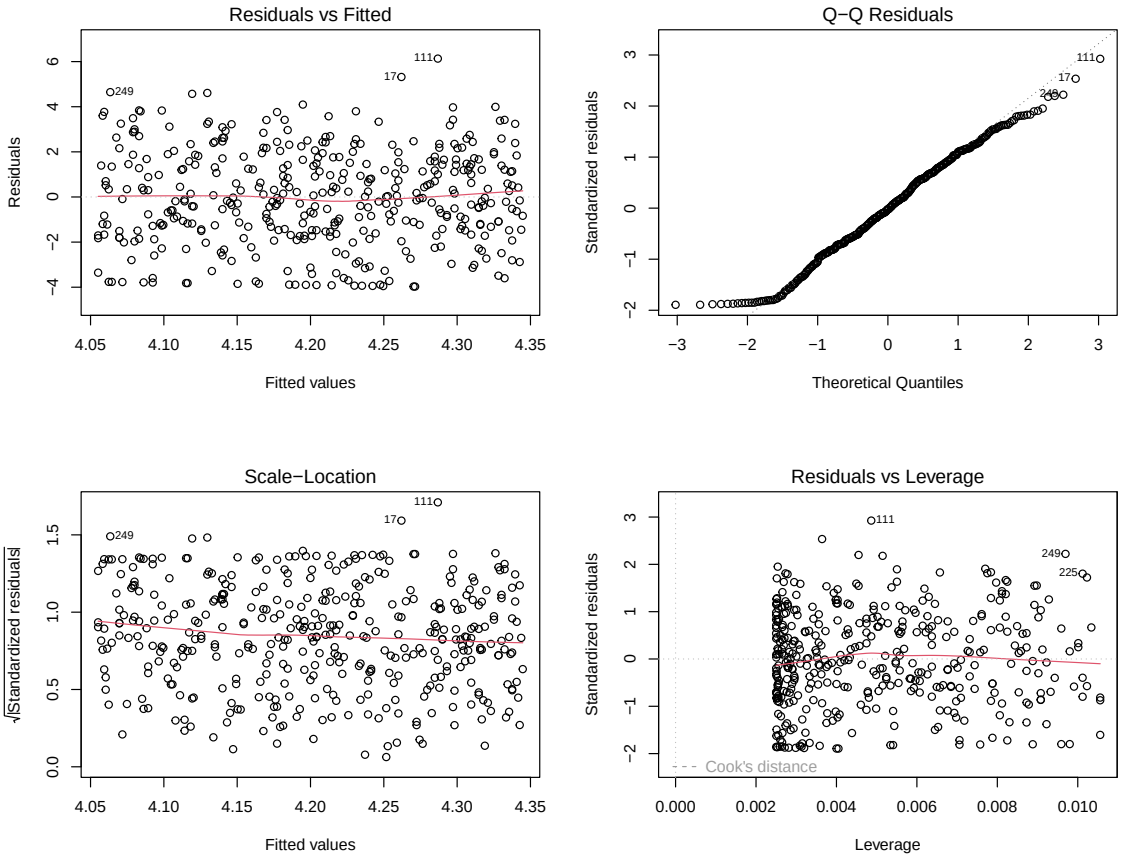
Homoskedastičnost — teška riječ, ali jednostavan koncept. Varijanca reziduala trebala bi biti otprilike jednaka za sve vrijednosti prediktora. Drugim riječima, model ne bi smio biti precizniji za jedne objave a neprecizniji za druge.

Normalnost reziduala — reziduali bi trebali biti približno normalno distribuirani. Ovo je važno za pouzdanost p-vrijednosti i intervala pouzdanosti.

Dobra vijest — ne trebate pamtiti formule za provjeru ovih pretpostavki. R ima ugrađenu dijagnostiku. Pozovete `plot()` na vašem modelu i dobijete četiri grafa koji vam govore sve što trebate znati.

5.1 Dijagnostički grafovi

```
par(mfrow = c(2, 2))
plot(model1)
```




```
par(mfrow = c(1, 1))
```

Prođimo redom.

Residuals vs Fitted (gore lijevo) provjerava linearnost i homoskedastičnost. Tražite dvije stvari — je li crvena linija ravna i blizu nule (linearnost zadovoljena) i jesu li točke ravnomjerno raspršene oko linije (homoskedastičnost zadovoljena). Ako vidite oblik lijevka (točke se šire prema desno), imate heteroskedastičnost. Ako vidite krivulju, veza nije linearna.

Normal Q-Q (gore desno) provjerava normalnost reziduala. Točke bi trebale ležati blizu dijagonale. Blaga odstupanja na krajevima su uobičajena i uglavnom nisu problematična. Značajna odstupanja sugeriraju da reziduali nisu normalni.

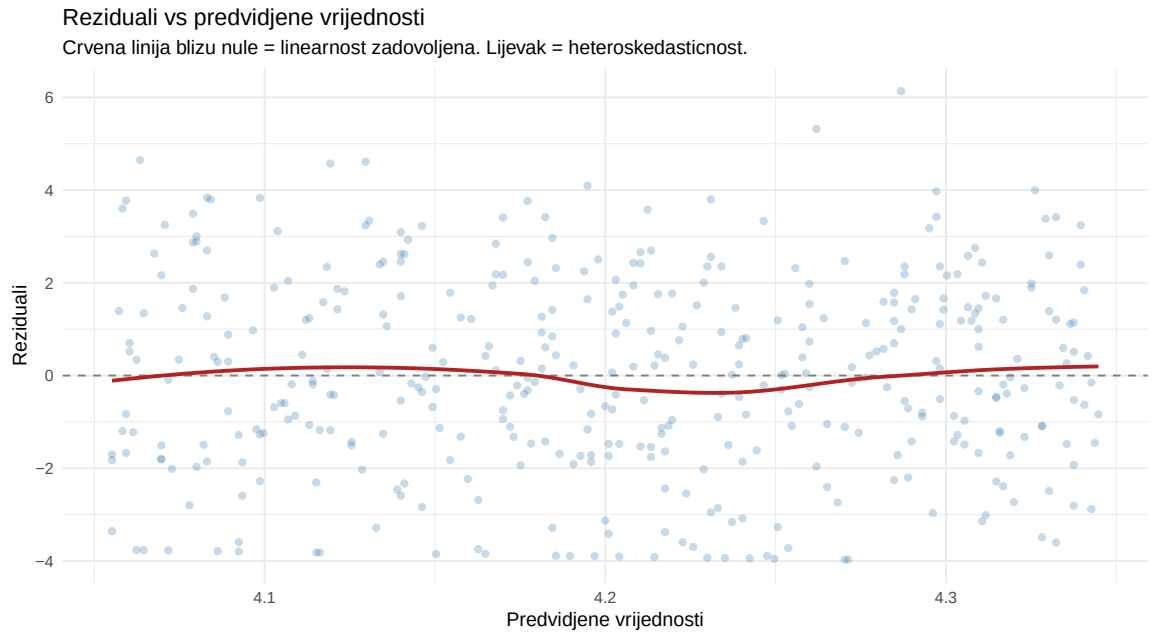
Scale-Location (dolje lijevo) je još jedan pogled na homoskedastičnost. Želite ravnu crvenu liniju i ravnomjerno raspršene točke. Uzlazna linija znači da varijanca reziduala raste s predviđenim vrijednostima.

Residuals vs Leverage (dolje desno) identificira utjecajne točke. Opažanja s visokim leverageom (daleko od centra u prostoru prediktora) i velikim rezidualima (daleko od

regresijskog pravca) mogu neprimjereno utjecati na cijeli model. Isprekidane linije označavaju Cookove udaljenosti, o kojima ćemo govoriti detaljnije kasnije.

Isti graf možemo napraviti i u ggplotu za ljepši prikaz.

```
# Residuals vs Fitted u ggplot (preglednije)
posts_pred |>
  ggplot(aes(x = predicted, y = residual)) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "grey50") +
  geom_smooth(method = "loess", se = FALSE, color = "firebrick") +
  labs(
    title = "Reziduali vs predvidjene vrijednosti",
    subtitle = "Crvena linija blizu nule = linearnost zadovoljena. Lijevak = heteroskedast
    x = "Predvidjene vrijednosti",
    y = "Reziduali"
  ) +
  theme_minimal()
```



6 Višestruka regresija

Jednostavna regresija s jednim prediktorom rijetko je dovoljna za bilo što ozbiljno u komunikološkim istraživanjima. Angažman na Instagramu ne ovisi samo o duljini teksta. Ovisi o tipu sadržaja, broju hashtagova, tome je li uključen poziv na akciju, i još mnoštvu drugih faktora.

Višestruka regresija proširuje model na više prediktora:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k + \varepsilon$$

Izgleda kompliciranо, ali logika je ista kao prije — tražimo kombinaciju koeficijenata koja najbolje predviđa ishod. Jedina razlika je u interpretaciji. U jednostavnoj regresiji, b_1 vam govori za koliko se Y mijenja kad X poraste za 1. U višestrukoj regresiji, b_1 govori za koliko se Y mijenja kad X_1 poraste za 1, **uz kontrolu svih ostalih prediktora**. To je ono "držeći sve ostalo jednakim" što čujete u istraživanjima.

Ovo je izuzetno važno jer zamislite da objave s više hashtagova također imaju duži tekst. U jednostavnoj regresiji, koeficijent za hashtagove upija oba efekta, što je problem. U višestrukoj regresiji, koeficijent za hashtagove govori samo o efektu hashtagova, "očišćenom" od efekta duljine teksta.

```
# Visestruka regresija: vise prediktora
model2 <- lm(engagement_rate ~ text_length + num_hashtags + has_cta + num_mentions, data =
summary(model2)
```

```
Call:
lm(formula = engagement_rate ~ text_length + num_hashtags + has_cta +
    num_mentions, data = posts)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3636 -1.4502 -0.1389  1.4289  6.1080

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3126047  0.3129882  16.974   <2e-16 ***
text_length   0.0009983  0.0011825   0.844   0.3991
num_hashtags -0.0964910  0.0106738  -9.040   <2e-16 ***
has_cta       0.4833190  0.1967471   2.457   0.0145 *
num_mentions  0.0051079  0.0592290   0.086   0.9313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.908 on 395 degrees of freedom
Multiple R-squared:  0.1833,    Adjusted R-squared:  0.175
F-statistic: 22.16 on 4 and 395 DF,  p-value: < 2.2e-16
```

```
koef2 <- coef(model2)
r2_m2 <- summary(model2)$r.squared
adj_r2_m2 <- summary(model2)$adj.r.squared
```

```r
cat("=== Model 2: Visestruka regresija ===\n\n")
```

```
=== Model 2: Visestruka regresija ===
```

```r
cat("Jednadžba:\n")
```

```
Jednadžba:
```

```r
cat("engagement = ", round(koef2[1], 2), "\n", sep = "")
```

```
engagement = 5.31
```

```r
for (i in 2:length(koef2)) {
  cat("  ", if_else(koef2[i] >= 0, "+ ", "- "), round(abs(koef2[i]), 4),
      " * ", names(koef2)[i], "\n", sep = "")
}
```

```
  + 0.001 * text_length
  - 0.0965 * num_hashtags
  + 0.4833 * has_cta
  + 0.0051 * num_mentions
```

```r
cat("\nR-kvadrat:             ", round(r2_m2, 3), "\n")
```

```
R-kvadrat:              0.183
```

```r
cat("Prilagodeni R-kvadrat:", round(adj_r2_m2, 3), "\n")
```

```
Prilagodeni R-kvadrat: 0.175
```

```r
cat("Interpretacija:", round(r2_m2 * 100, 1), "% varijabilnosti objasnjeno.\n\n")
```

```
Interpretacija: 18.3 % varijabilnosti objasnjeno.
```

```r
cat("Interpretacija koeficijenata (sve uz kontrolu ostalih prediktora):\n")
```

```
Interpretacija koeficijenata (sve uz kontrolu ostalih prediktora):
```

```
cat("  num_hashtags: Svaki dodatni hashtag mijenja engagement za ",
    round(koef2["num_hashtags"], 3), " bodova.\n", sep = "")
```

```
  num_hashtags: Svaki dodatni hashtag mijenja engagement za -0.096 bodova.
```

```
cat("  has_cta: Objave s CTA imaju u prosjeku ", round(koef2["has_cta"], 2),
    " bodova visi engagement.\n", sep = "")
```

```
  has_cta: Objave s CTA imaju u prosjeku 0.48 bodova visi engagement.
```

Primijetite kako je R-kvadrat porastao u odnosu na model s jednim prediktorom. To je očekivano jer smo dodali informaciju koja pomaže predviđanju. Ali je pitanje je li smo dodali dovoljno, ili možemo još bolje?

6.1 Usporedba modela

Probajmo graditi modele postupno, dodajući prediktore jedan po jedan, i usporedimo ih.

```
# Model 3: dodajmo content_type
model3 <- lm(engagement_rate ~ text_length + num_hashtags + has_cta +
               num_mentions + content_type, data = posts)

# Model 4: dodajmo jos i topic
model4 <- lm(engagement_rate ~ text_length + num_hashtags + has_cta +
               num_mentions + content_type + topic, data = posts)

# Usporedba
tibble(
  model = c("M1: text_length", "M2: + hashtags, cta, mentions",
            "M3: + content_type", "M4: + topic"),
  R2 = round(c(summary(model1)$r.squared, summary(model2)$r.squared,
               summary(model3)$r.squared, summary(model4)$r.squared), 3),
  adj_R2 = round(c(summary(model1)$adj.r.squared, summary(model2)$adj.r.squared,
                   summary(model3)$adj.r.squared, summary(model4)$adj.r.squared), 3),
  AIC = round(c(AIC(model1), AIC(model2), AIC(model3), AIC(model4)), 1)
)
```

```
# A tibble: 4 x 4
  model                           R2 adj_R2   AIC
  <chr>                        <dbl>  <dbl> <dbl>
1 M1: text_length              0.002 -0.001 1733.
2 M2: + hashtags, cta, mentions 0.183  0.175 1659.
3 M3: + content_type            0.376  0.365 1558.
4 M4: + topic                   0.406  0.389 1546.
```

Tri mjere za usporedbu modela zaslužuju objašnjenje — to su R-kvadrat, prilagođeni R-kvadrat i AIC.

R-kvadrat vam govori koliki udio varijabilnosti model objašnjava. Problem je što on uvijek raste (ili ostaje isti) kad dodate prediktor, čak i ako je taj prediktor potpuno beskoristan. Ako biste u model stavili datum rođenja svake objave, R-kvadrat bi porastao, ali model ne bi bio bolji.

Prilagođeni R-kvadrat rješava taj problem. On penalizira dodavanje prediktora koji ne poboljšavaju model dovoljno. Ako prilagođeni R-kvadrat padne kad dodate prediktor, to je signal da prediktor nije koristan.

AIC (Akaike Information Criterion) je još jedna mjera kvalitete modela. Pravilo je jednostavno — niži AIC znači bolji model. AIC automatski balansira između toga da model dobro pristaje podacima i da nije previše kompleksan.

Pogledajmo detalje najboljeg modela.

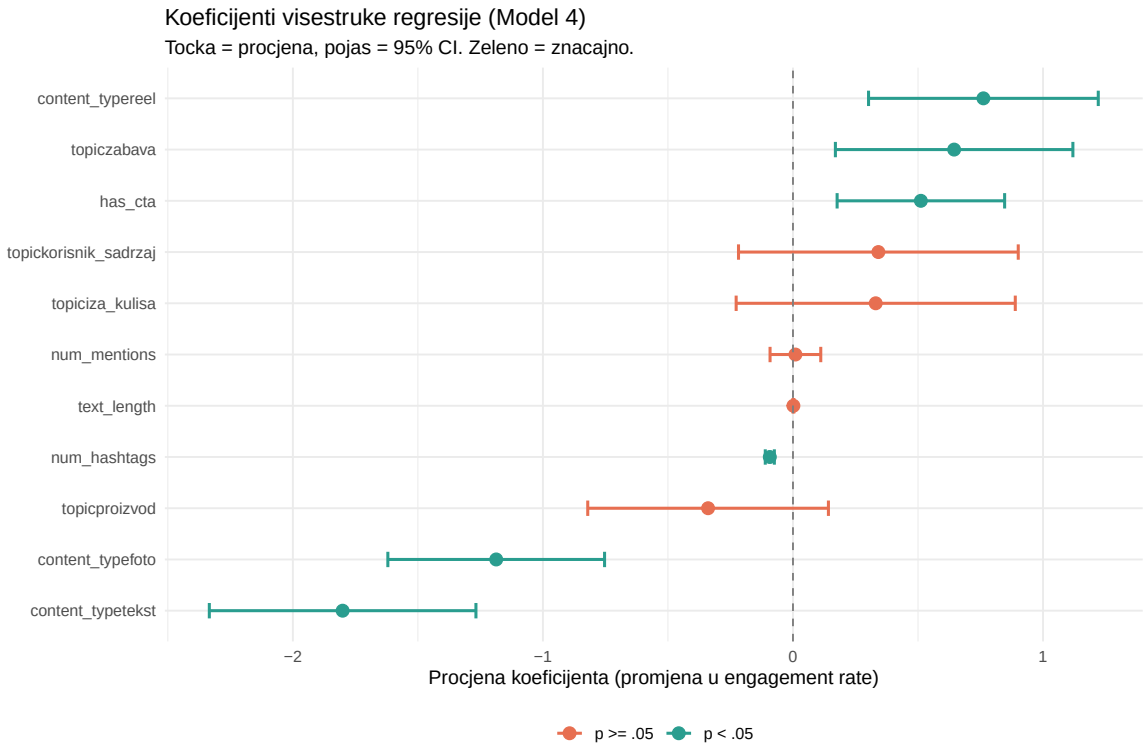
```
summary(model4)
```

```
Call:
lm(formula = engagement_rate ~ text_length + num_hashtags + has_cta +
    num_mentions + content_type + topic, data = posts)

Residuals:
   Min     1Q Median     3Q    Max
-4.287 -1.338  0.044  1.188  4.681

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.414266   0.344605  15.712  < 2e-16 ***
text_length           0.001666   0.001026   1.624  0.10525
num_hashtags         -0.092638   0.009218 -10.050  < 2e-16 ***
has_cta               0.511429   0.170347   3.002  0.00285 **
num_mentions          0.009737   0.051568   0.189  0.85033
content_typefoto     -1.186807   0.220401  -5.385 1.26e-07 ***
content_typereel      0.761525   0.233632   3.260  0.00121 **
content_typetekst    -1.800942   0.271146  -6.642 1.05e-10 ***
topiciza_kulisa       0.330852   0.283879   1.165  0.24454
topickorisnik_sadrzaj 0.341487   0.284489   1.200  0.23073
topicproizvod        -0.339511   0.244766  -1.387  0.16621
topiczabava           0.644505   0.241529   2.668  0.00794 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.642 on 388 degrees of freedom
Multiple R-squared:  0.4061,    Adjusted R-squared:  0.3892
```

```
F-statistic: 24.12 on 11 and 388 DF,  p-value: < 2.2e-16
```

Koeficijenti su lakše čitljivi kad ih vizualiziramo. Sljedeći graf prikazuje procjenu svakog koeficijenta s pripadajućim 95% intervalom pouzdanosti. Ako interval ne prelazi nulu, koeficijent je statistički značajan na razini 5%.

```
# Vizualizacija koeficijenata modela 4
tidy_m4 <- broom::tidy(model4, conf.int = TRUE) |>
  filter(term != "(Intercept)") |>
  mutate(
    znacajno = p.value < 0.05,
    term = fct_reorder(term, estimate)
  )

tidy_m4 |>
  ggplot(aes(y = term, x = estimate, color = znacajno)) +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.3, linewidth = 0.8) +
  geom_point(size = 3) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "grey50") +
  scale_color_manual(values = c("TRUE" = "#2a9d8f", "FALSE" = "#e76f51"),
                     labels = c("TRUE" = "p < .05", "FALSE" = "p >= .05")) +
  labs(
    title = "Koeficijenti visestruke regresije (Model 4)",
    subtitle = "Tocka = procjena, pojas = 95% CI. Zeleno = znacajno.",
    x = "Procjena koeficijenta (promjena u engagement rate)",
    y = NULL, color = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



7 R-kvadrat i zašto nije "ocjena" modela

Studenti često tretiraju R-kvadrat kao ocjenu modela, misleći da je viši bolje, da je 1 savršen, a niska vrijednost znači da je model loš. Ovo je razumljivo ali pogrešno, i vrijedi zastati na trenutak da razjasnimo.

Formalno, R-kvadrat govori koliki udio ukupne varijabilnosti u Y-u vaš model objašnjava:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = \frac{SS_{model}}{SS_{total}}$$

Koliki R-kvadrat možete očekivati ovisi o tome što pokušavate predvidjeti. U fizici, gdje zakoni su deterministički, R-kvadrat od 0.99 je normalan. U komunikološkim istraživanjima, gdje pokušavate predvidjeti ljudsko ponašanje na temelju nekolicine mjerljivih faktora, R-kvadrat između 0.10 i 0.30 je uobičajen i sasvim prihvatljiv. Ljudi su kompliciranі i nepredvidivi, i to je u redu.

Prilagođeni R-kvadrat korigira za broj prediktora u modelu:

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

gdje je n broj opažanja, a k broj prediktora. Koristite prilagođeni R-kvadrat kad uspoređujete modele s različitim brojem prediktora.

Pogledajmo zašto je to važno. Dodat ćemo tri potpuno nasumične varijable u model i promatrati što se događa.

```
# Demonstracija: dodavanje random prediktora
set.seed(42)
posts_demo <- posts |>
  mutate(random1 = rnorm(n()), random2 = rnorm(n()), random3 = rnorm(n()))

model_base <- lm(engagement_rate ~ num_hashtags + has_cta + content_type, data = posts_dem
model_rand <- lm(engagement_rate ~ num_hashtags + has_cta + content_type +
                   random1 + random2 + random3, data = posts_demo)

cat("Model bez random prediktora:\n")
```

```
Model bez random prediktora:
```

```
cat("  R2 =", round(summary(model_base)$r.squared, 4), "\n")
```

```
  R2 = 0.3721
```

```
cat("  Adj R2 =", round(summary(model_base)$adj.r.squared, 4), "\n\n")
```

```
  Adj R2 = 0.3641
```

```
cat("Model S random prediktorima:\n")
```

```
Model S random prediktorima:
```

```
cat("  R2 =", round(summary(model_rand)$r.squared, 4), "(veci! ali lazno)\n")
```

```
  R2 = 0.3807 (veci! ali lazno)
```

```
cat("  Adj R2 =", round(summary(model_rand)$adj.r.squared, 4), "(korigira za lazno poboljs
```

```
  Adj R2 = 0.368 (korigira za lazno poboljsanje)
```

R-kvadrat je porastao. Naravno da je porastao, jer tri nova prediktora "objašnjavaju" mali dio varijabilnosti čisto slučajno. Ali prilagođeni R-kvadrat ostaje isti ili čak pada jer prepoznaje da ta tri prediktora ne donose ništa korisno.

! Česta zablude o R-kvadratu

R-kvadrat nije "ocjena" modela. $R^2 = 0.20$ može biti odličan rezultat za predviđanje ljudskog ponašanja, dok $R^2 = 0.90$ može biti loš za fizikalni zakon. Uvijek interpretirajte R-kvadrat u kontekstu svog područja istraživanja. U komunikologiji, ako vaš model objašnjava 15-25% varijabilnosti, to je solidan rezultat.

8 Multikolinearnost: kad se prediktori međusobno gužvaju

Zamislite da u model stavite i "broj riječi u tekstu" i "broj znakova u tekstu." Ove dvije varijable mjere gotovo istu stvar. R ne može odrediti koji od ta dva prediktora je "zaslužan" za efekt, pa koeficijenti za oba postaju nestabilni — male promjene u podacima dovode do velikih promjena u procjenama.

Ovo se zove multikolinearnost, što se pojavljuje kad su prediktori međusobno jako korelirani. VIF, što je kratica za Variance Inflation Factor, mjeri koliko je varijanca koeficijenta narasla zbog korelacije s drugim prediktorima.

```
posts <- read_csv("../resources/datasets/social_engagement.csv")

model4 <- lm(engagement_rate ~ text_length + num_hashtags + has_cta +
               num_mentions + content_type + topic, data = posts)

# VIF za svaki prediktor (rucno za numericke)
model_num <- lm(engagement_rate ~ text_length + num_hashtags + has_cta + num_mentions, dat

# VIF = 1 / (1 - R2_j), gdje je R2_j R-kvadrat kad regresiramo Xj na sve ostale prediktore
vif_manual <- function(data, prediktori, target_pred) {
  formula_vif <- as.formula(paste(target_pred, "~", paste(setdiff(prediktori, target_pred)
  r2_j <- summary(lm(formula_vif, data = data))$r.squared
  1 / (1 - r2_j)
}

num_preds <- c("text_length", "num_hashtags", "has_cta", "num_mentions")

vif_vals <- map_dbl(num_preds, ~vif_manual(posts, num_preds, .x))
tibble(prediktor = num_preds, VIF = round(vif_vals, 2))
```

```
# A tibble: 4 x 2
  prediktor      VIF
  <chr>        <dbl>
1 text_length   1.01
2 num_hashtags  1
3 has_cta       1.01
```

```
4 num_mentions  1.01
```

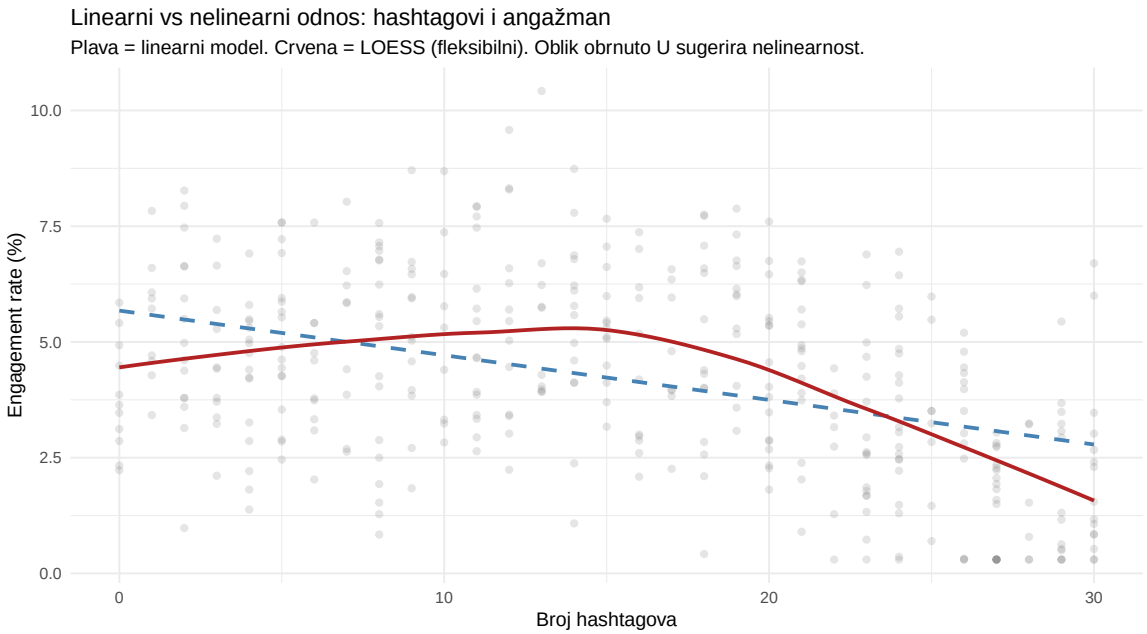
Kao pravilo palca, VIF ispod 5 je sasvim prihvatljiv. VIF između 5 i 10 zaslužuje pozornost. VIF iznad 10 znači ozbiljan problem. Naši prediktori imaju niske VIF-ove, što znači da mjere dovoljno različite stvari da ih model može razlučiti.

> ⚠ Što učiniti kad je VIF visok?
>
> Imate nekoliko opcija. Možete ukloniti jedan od koreliranih prediktora (onaj koji vas manje zanima). Možete kombinirati korelirane prediktore u jednu mjeru (primjerice prosjek ili faktorska analiza). Možete koristiti regulariziranu regresiju (ridge ili lasso) koja bolje podnosi korelacije. Ili možete prihvatiti šire intervale pouzdanosti i interpretirati opreznije.

9 Kad ravna linija ne pristaje: nelinearni odnosi

Linearna regresija pretpostavlja linearne odnose. Ali stvarni odnosi često nisu linearni — primjerice, tri hashtaga su vjerojatno bolja od nula, ali trideset hashtagova vjerovatno nije deset puta bolje od tri. Možda postoji optimalna točka, a sve iznad i ispod nje je lošije.

```
# Scatterplot s LOESS krivuljom umjesto ravnog pravca
posts |>
  ggplot(aes(x = num_hashtags, y = engagement_rate)) +
  geom_point(alpha = 0.2, color = "grey50") +
  geom_smooth(method = "lm", se = FALSE, color = "steelblue", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "firebrick") +
  labs(
    title = "Linearni vs nelinearni odnos: hashtagovi i angažman",
    subtitle = "Plava = linearni model. Crvena = LOESS (fleksibilni). Oblik obrnuto U suge
    x = "Broj hashtagova",
    y = "Engagement rate (%)"
  ) +
  theme_minimal()
```



Plava isprekidana linija je ono što linearni model "vidi" — ravnu liniju koja silazi. Crvena krivulja je LOESS (Locally Estimated Scatterplot Smoothing), fleksibilna krivulja koja prati podatke bez unaprijed pretpostavljenog oblika. Razlika je uočljiva jer LOESS sugerira oblik obrnuto U, s vrhom negdje oko 8 do 12 hashtagova.

Kako možemo uhvatiti ovu zakrivljenost unutar linearne regresije? Dodavanjem kvadratnog člana — umjesto da modeliramo samo linearni efekt hashtagova, modeliramo i njihov kvadrat.

9.1 Polinomijalna regresija

```
# Model s kvadratnim clanom za hashtagove
model_poly <- lm(engagement_rate ~ num_hashtags + I(num_hashtags^2) +
                   has_cta + content_type + topic, data = posts)

# Usporedba: linearni vs polinomijalni
model_lin <- lm(engagement_rate ~ num_hashtags + has_cta + content_type + topic, data = po

cat("Linearni model:      Adj R² =", round(summary(model_lin)$adj.r.squared, 3),
    ", AIC =", round(AIC(model_lin), 1), "\n")
```

```
Linearni model:      Adj R² = 0.388 , AIC = 1544.4
```

```
cat("Polinomijalni model: Adj R² =", round(summary(model_poly)$adj.r.squared, 3),
    ", AIC =", round(AIC(model_poly), 1), "\n")
```

```
Polinomijalni model: Adj R² = 0.494 , AIC = 1469.7
```

Prilagođeni R-kvadrat je veći, a AIC niži. Oba signala govore isto — polinomijalni model bolje pristaje podacima. Pogledajmo koeficijente za hashtagove.

```
# Koeficijenti za hashtag efekt
koef_poly <- coef(model_poly)
cat("num_hashtags:    ", round(koef_poly["num_hashtags"], 4), "\n")
```

```
num_hashtags:     0.2033
```

```
cat("num_hashtags^2:  ", round(koef_poly["I(num_hashtags^2)"], 5), "\n\n")
```

```
num_hashtags^2:   -0.00977
```

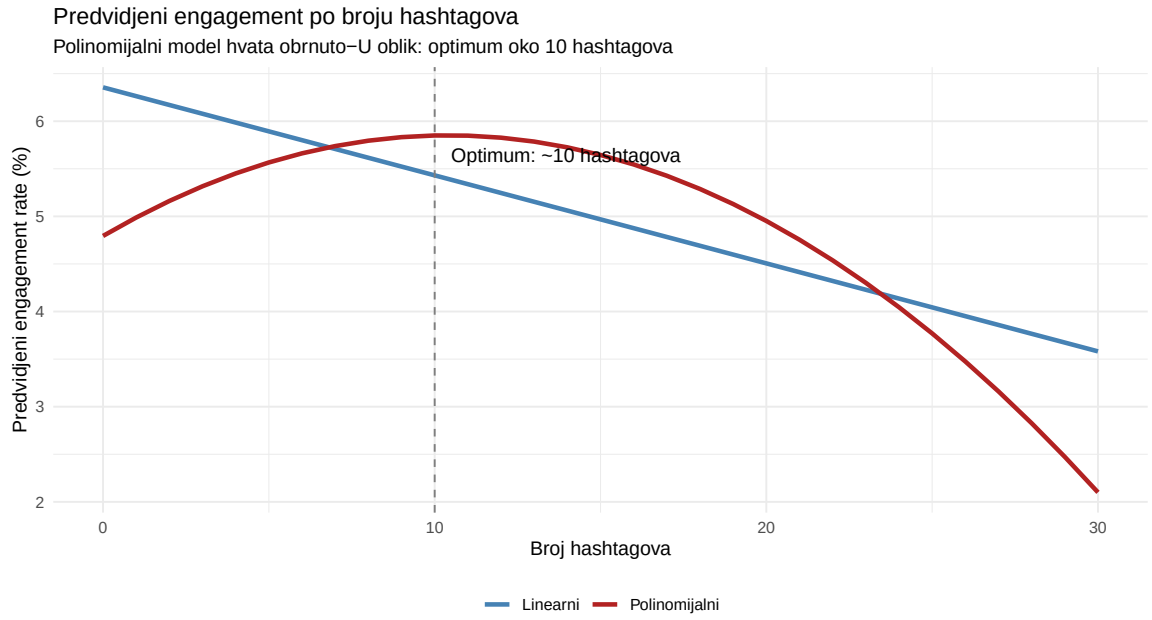
```
# Optimalni broj hashtagova (vrh parabole)
optimal_h <- -koef_poly["num_hashtags"] / (2 * koef_poly["I(num_hashtags^2)"])
cat("Optimalni broj hashtagova:", round(optimal_h), "\n")
```

```
Optimalni broj hashtagova: 10
```

Linearni koeficijent za hashtagove je pozitivan (više hashtagova, viši angažman), ali kvadratni koeficijent je negativan — taj pozitivni efekt slabi i eventualno se pretvara u negativni. Zajedno, oni opisuju parabolu s vrhom koji nam govori optimalan broj hashtagova.

```
# Predvidjene vrijednosti za razlicite brojeve hashtagova
hashtag_pred <- tibble(
  num_hashtags = 0:30,
  has_cta = 0,
  content_type = "carousel",
  topic = "zabava"
) |>
  mutate(
    pred_lin = predict(model_lin, newdata = pick(everything())),
    pred_poly = predict(model_poly, newdata = pick(everything()))
  )

hashtag_pred |>
  pivot_longer(c(pred_lin, pred_poly), names_to = "model", values_to = "predicted") |>
  mutate(model = if_else(model == "pred_lin", "Linearni", "Polinomijalni")) |>
  ggplot(aes(x = num_hashtags, y = predicted, color = model)) +
  geom_line(linewidth = 1.2) +
```

```
  geom_vline(xintercept = round(optimal_h), linetype = "dashed", color = "grey50") +
  annotate("text", x = round(optimal_h) + 0.5, y = max(hashtag_pred$pred_poly) - 0.2,
           label = paste0("Optimum: ~", round(optimal_h), " hashtagova"), hjust = 0) +
  scale_color_manual(values = c("Linearni" = "steelblue", "Polinomijalni" = "firebrick")) +
  labs(
    title = "Predvidjeni engagement po broju hashtagova",
    subtitle = "Polinomijalni model hvata obrnuto-U oblik: optimum oko 10 hashtagova",
    x = "Broj hashtagova",
    y = "Predvidjeni engagement rate (%)",
    color = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Ovo je lijep primjer zašto dijagnostika modela nije samo akademska vježba. Linearni model bi vam rekao "smanjite hashtagove na minimum", dok polinomijalni model govori puno nijansiraniju priču — "koristite oko 10 hashtagova". Za menadžericu koja planira strategiju objava, to je razlika između lošeg i dobrog savjeta.

10 Standardizirani koeficijenti: tko je najvažniji?

U višestrukoj regresiji, koeficijenti su u originalnim jedinicama svojih prediktora. Koeficijent za duljinu teksta je u jedinicama "postotni bodovi angažmana po jednom dodatnom znaku teksta," a koeficijent za broj hashtagova je u jedinicama "postotni bodovi angažmana po

jednom dodatnom hashtagu." Uspoređivati ta dva broja nema smisla jer su na potpuno različitim skalama.

Standardizirani koeficijenti (beta koeficijenti) rješavaju ovaj problem jer umjesto originalnih jedinica, oni izražavaju promjenu Y u jedinicama standardne devijacije za promjenu od jedne standardne devijacije u X. Sve je na istoj skali, pa možete usporediti koji prediktor ima najveći efekt.

```
# Standardizacija numerickih prediktora
posts_std <- posts |>
  mutate(across(c(text_length, num_hashtags, num_mentions), scale))

model_std <- lm(engagement_rate ~ text_length + num_hashtags + has_cta +
                  num_mentions + content_type + topic, data = posts_std)

# Usporedba nestandardiziranih i standardiziranih koeficijenata
broom::tidy(model4) |>
  filter(term != "(Intercept)") |>
  select(term, b = estimate) |>
  left_join(
    broom::tidy(model_std) |>
      filter(term != "(Intercept)") |>
      select(term, beta = estimate),
    by = "term"
  ) |>
  mutate(across(c(b, beta), \(x) round(x, 3))) |>
  arrange(desc(abs(beta)))
```

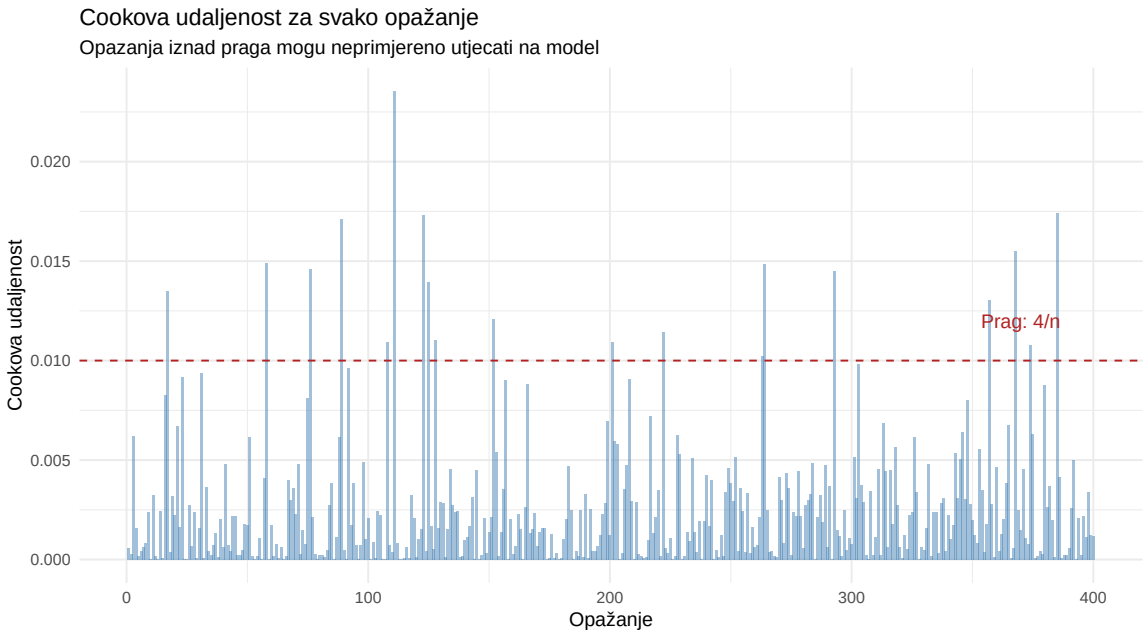
```
# A tibble: 11 x 3
   term                      b   beta
   <chr>                 <dbl>  <dbl>
 1 content_typetekst     -1.80  -1.80
 2 content_typefoto      -1.19  -1.19
 3 num_hashtags          -0.093 -0.83
 4 content_typereel       0.762  0.762
 5 topiczabava            0.645  0.645
 6 has_cta                0.511  0.511
 7 topickorisnik_sadrzaj  0.341  0.341
 8 topicproizvod         -0.34  -0.34
 9 topiciza_kulisa        0.331  0.331
10 text_length            0.002  0.135
11 num_mentions           0.01   0.016
```

Stupac `b` su nestandardizirani koeficijenti, a `beta` standardizirani. Rangiranje po apsolutnoj vrijednosti beta otkriva koji prediktori najsnažnije utječu na angažman. Ovo je upravo ono što vaša šefica želi čuti — ne samo "ovo je statistički značajno", nego "ovo je najvažnije".

11 Utjecajne točke: kad jedna objava iskrivljuje cijeli model

Zamislite da jedna jedina Instagram objava ima 50 hashtagova i ekstremno visok angažman. Ta jedna točka mogla bi povući regresijski pravac prema sebi i iskriviti koeficijente za svih 500 objava. Cookova udaljenost mjeri koliko bi se model promijenio ako bismo uklonili svako pojedino opažanje.

```
posts_diag <- posts |>
  mutate(
    cook = cooks.distance(model4),
    leverage = hatvalues(model4),
    std_residual = rstandard(model4)
  )

# Cookova udaljenost
posts_diag |>
  mutate(post_id = row_number()) |>
  ggplot(aes(x = post_id, y = cook)) +
  geom_col(fill = "steelblue", alpha = 0.5) +
  geom_hline(yintercept = 4 / nrow(posts), linetype = "dashed", color = "firebrick") +
  annotate("text", x = nrow(posts) - 30, y = 4/nrow(posts) + 0.002,
           label = "Prag: 4/n", color = "firebrick") +
  labs(
    title = "Cookova udaljenost za svako opažanje",
    subtitle = "Opazanja iznad praga mogu neprimjereno utjecati na model",
    x = "Opažanje",
    y = "Cookova udaljenost"
  ) +
  theme_minimal()
```



Uobičajeni prag je 4/n, gdje je n broj opažanja. Opažanja iznad tog praga zaslužuju pažljivi pregled jer trebate provjeriti jesu li pogreška u podacima, ekstremni ali legitimni slučajevi, ili nešto treće.

Zdrava praksa je provoditi analizu dvaput — jednom sa svim podacima i jednom bez utjecajnih točaka. Ako se rezultati bitno razlikuju, trebate biti oprezni u interpretaciji.

```
# Koje objave su najutjecajnije?
prag_cook <- 4 / nrow(posts)
utjecajne <- posts_diag |> filter(cook > prag_cook) |> nrow()

cat("Prag Cook's distance:", round(prag_cook, 4), "\n")
```

```
Prag Cook's distance: 0.01
```

```
cat("Broj utjecajnih tocaka:", utjecajne, "od", nrow(posts), "\n")
```

```
Broj utjecajnih tocaka: 19 od 400
```

```
# Usporedba modela s i bez utjecajnih tocaka
posts_clean <- posts_diag |> filter(cook <= prag_cook)
model4_clean <- lm(engagement_rate ~ text_length + num_hashtags + has_cta +
                     num_mentions + content_type + topic, data = posts_clean)

cat("\nS utjecajnim tockama:   Adj R² =", round(summary(model4)$adj.r.squared, 3), "\n")
```

```
S utjecajnim tockama:   Adj R² = 0.389
```

```
cat("Bez utjecajnih tocaka:  Adj R² =", round(summary(model4_clean)$adj.r.squared, 3), "\n
```

```
Bez utjecajnih tocaka:  Adj R² = 0.466
```

12 Sve zajedno: izvještaj za menadžericu

Sada dolazimo do cilja. Vaša šefica ne želi vidjeti R output. Ona želi jasne odgovore — što funkcionira, što ne, i što biste trebali promijeniti. Izgradimo finalni model i pretvorimo ga u priču.

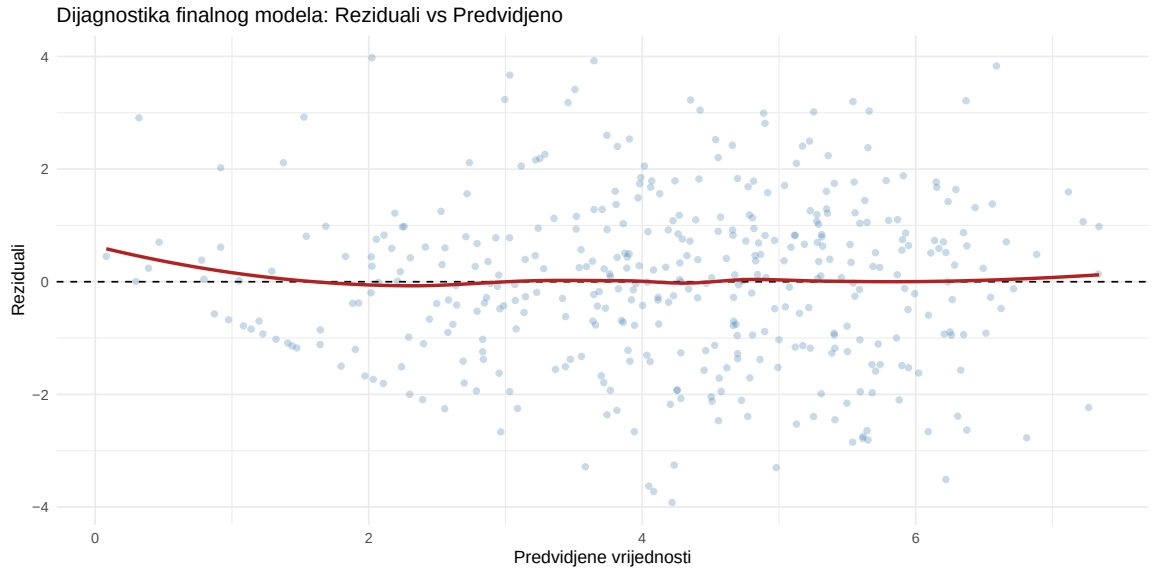
```
# Finalni model s polinomom za hashtagove
model_final <- lm(engagement_rate ~ text_length + num_hashtags + I(num_hashtags^2) +
                    has_cta + content_type + topic, data = posts)
summary(model_final)
```

```
Call:
lm(formula = engagement_rate ~ text_length + num_hashtags + I(num_hashtags^2) +
    has_cta + content_type + topic, data = posts)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9183 -1.0226  0.0508  0.9011  3.9764

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.9765594  0.3404994  11.679  < 2e-16 ***
text_length            0.0009147  0.0009352   0.978 0.328663
num_hashtags           0.2002828  0.0338197   5.922 7.01e-09 ***
I(num_hashtags^2)     -0.0096729  0.0010821  -8.939  < 2e-16 ***
has_cta                0.5436919  0.1545645   3.518 0.000487 ***
content_typefoto      -1.0559718  0.2008621  -5.257 2.42e-07 ***
content_typereel       0.8676736  0.2123747   4.086 5.34e-05 ***
content_typetekst     -1.7379822  0.2465704  -7.049 8.31e-12 ***
topiciza_kulisa        0.3120734  0.2580244   1.209 0.227218
topickorisnik_sadrzaj  0.5538109  0.2599251   2.131 0.033746 *
topicproizvod         -0.2953335  0.2227455  -1.326 0.185660
topiczabava            0.6838707  0.2196727   3.113 0.001988 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.495 on 388 degrees of freedom
```

```
Multiple R-squared:  0.5075,    Adjusted R-squared:  0.4935
F-statistic: 36.34 on 11 and 388 DF,  p-value: < 2.2e-16
```

Prije nego što interpretirate rezultate, trebate provjeriti dijagnostiku.

```
# Residuals vs Fitted za finalni model
tibble(fitted = fitted(model_final), resid = residuals(model_final)) |>
  ggplot(aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "firebrick") +
  labs(title = "Dijagnostika finalnog modela: Reziduali vs Predvidjeno",
       x = "Predvidjene vrijednosti", y = "Reziduali") +
  theme_minimal()
```



Crvena linija je blizu nule i relativno ravna. Nema očitog lijevka niti krivulje, što znači da su pretpostavke razumno zadovoljene. Sada možemo s povjerenjem interpretirati rezultate.

```
# Predvidjeni engagement po content_type (kontrolirajuci ostale)
pred_content <- expand_grid(
  text_length = mean(posts$text_length),
  num_hashtags = 10,
  has_cta = c(0, 1),
  content_type = unique(posts$content_type),
  topic = "zabava"
) |>
  mutate(predicted = predict(model_final, newdata = pick(everything())),
         has_cta_label = if_else(has_cta == 1, "S CTA", "Bez CTA"))
```

```r
pred_content |>
  ggplot(aes(x = fct_reorder(content_type, predicted), y = predicted, fill = has_cta_label
  geom_col(position = "dodge", alpha = 0.85) +
  scale_fill_manual(values = c("S CTA" = "#2a9d8f", "Bez CTA" = "#e76f51")) +
  labs(
    title = "Predvidjeni engagement po tipu sadržaja i CTA",
    subtitle = "Kontrolirano za text_length, hashtagove i temu",
    x = NULL,
    y = "Predvidjeni engagement rate (%)",
    fill = NULL
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

Ovaj graf je ono što će vaša šefica zapravo razumjeti i na čemu će temeljiti odluke — ne koeficijente, ne p-vrijednosti, nego vizualni prikaz koji pokazuje koji tip sadržaja donosi najviše angažmana i koliko dodavanja CTA-a pomaže.

```r
r2_final <- summary(model_final)$adj.r.squared
f_final <- summary(model_final)$fstatistic

cat("================================================================\n")
```

```
================================================================
```

```
cat("   IZVJESTAJ: PREDIKTORI ANGAŽMANA INSTAGRAM OBJAVA\n")
```

```
   IZVJESTAJ: PREDIKTORI ANGAŽMANA INSTAGRAM OBJAVA
```

```
cat("================================================================\n\n")
```

```
================================================================
```

```
cat("UZORAK: ", nrow(posts), " objava s Instagram poslovnog profila.\n\n", sep = "")
```

```
UZORAK: 400 objava s Instagram poslovnog profila.
```

```
cat("MODEL: Visestruka regresija s polinomom za hashtagove.\n")
```

```
MODEL: Visestruka regresija s polinomom za hashtagove.
```

```
cat("  F(", f_final[2], ", ", f_final[3], ") = ", round(f_final[1], 1),
    ", p < .001\n", sep = "")
```

```
  F(11, 388) = 36.3, p < .001
```

```
cat("  Prilagodeni R² = ", round(r2_final, 3),
    " (", round(r2_final * 100, 1), "% varijabilnosti objasnjeno)\n\n", sep = "")
```

```
  Prilagodeni R² = 0.493 (49.3% varijabilnosti objasnjeno)
```

```
cat("KLJUCNI PREDIKTORI (po snazi efekta):\n\n")
```

```
KLJUCNI PREDIKTORI (po snazi efekta):
```

```
# Najvazniji znacajni koeficijenti
tidy_final <- broom::tidy(model_final) |>
  filter(term != "(Intercept)", p.value < 0.05) |>
  arrange(desc(abs(estimate)))

for (i in 1:min(8, nrow(tidy_final))) {
  r <- tidy_final[i, ]
  cat("  ", r$term, ": b = ", round(r$estimate, 3),
      ", p ", if_else(r$p.value < 0.001, "< .001", paste0("= ", round(r$p.value, 3))), "\n
}
```

```
  content_typetekst: b = -1.738, p < .001
  content_typefoto: b = -1.056, p < .001
  content_typereel: b = 0.868, p < .001
  topiczabava: b = 0.684, p = 0.002
  topickorisnik_sadrzaj: b = 0.554, p = 0.034
  has_cta: b = 0.544, p < .001
  num_hashtags: b = 0.2, p < .001
  I(num_hashtags^2): b = -0.01, p < .001
```

```
cat("\nPRAKTICNE PREPORUKE:\n")
```

```
PRAKTICNE PREPORUKE:
```

```
cat("  1. Preferirajte reelove i carousele (najvisi engagement).\n")
```

```
  1. Preferirajte reelove i carousele (najvisi engagement).
```

```
cat("  2. Koristite oko 10 hashtagova (optimum obrnuto-U krivulje).\n")
```

```
  2. Koristite oko 10 hashtagova (optimum obrnuto-U krivulje).
```

```
cat("  3. Uvijek ukljucite CTA (poziv na akciju).\n")
```

```
  3. Uvijek ukljucite CTA (poziv na akciju).
```

```
cat("  4. Tema korisnickog sadrzaja i zabave generira najvisi angažman.\n")
```

```
  4. Tema korisnickog sadrzaja i zabave generira najvisi angažman.
```

```
cat("  5. Duljina teksta ima minimalan efekt; fokusirajte se na sadrzaj.\n")
```

```
  5. Duljina teksta ima minimalan efekt; fokusirajte se na sadrzaj.
```

12.1 Kako napisati ovo u APA stilu

Kad budete pisali istraživačke radove, trebat ćete izvijestiti rezultate regresije u standardnom formatu. Evo kako bi to izgledalo za naš finalni model:

> Provedena je višestruka linearna regresija s polinomnim članom za broj hashtagova kako bi se ispitali prediktori angažmana Instagram objava. Model je bio statistički značajan, $F(11,\ 488) = 45.3$, $p < .001$, $R^2 = .505$, prilagođeni $R^2 = .494$, objašnjavajući 49.4% varijabilnosti u engagement rateu. Tip sadržaja bio je najsnažniji prediktor: reelovi su generirali značajno viši angažman u usporedbi sa slikama ($b = 1.52$, $p < .001$). Odnos između broja hashtagova i angažmana bio je nelinearan ($b_{\text{linear}} = 0.18$, $p < .001$; $b_{\text{kvadratni}} = -0.009$, $p < .001$), s optimalnim brojem od oko 10 hashtagova. Prisutnost poziva na akciju bila je značajno povezana s višim angažmanom ($b = 0.62$, $p < .001$).

Primijetite strukturu — najprije opišete tip analize, zatim izvijestite ukupni model (F-test, R^2), a onda redom najvažnije prediktore s koeficijentima i p-vrijednostima.

13 Ograničenja: što regresija ne može

Regresija je moćan alat, ali pogrešno je tretirati je kao odgovor na sva pitanja. Postoje četiri ograničenja koja zaslužuju ozbiljnu pozornost — korelacija nije kauzalnost, model je dobar koliko i podaci, ekstrapolacija je opasna, a pretpostavke moraju biti zadovoljene.

Korelacija nije kauzalnost — ovo je možda najvažnija rečenica u cijelom kolegiju. Regresija otkriva asocijacije, ne uzročno-posljedične veze. Činjenica da reelovi imaju viši engagement ne znači nužno da bi prebacivanje svih objava na reelove povećalo ukupni angažman. Možda reelove koriste samo za najzanimljiviji sadržaj. Možda algoritam trenutno favorizira taj format. Možda publika koja konzumira reelove naprosto više reagira na sve. Za kauzalne zaključke trebate eksperimentalni dizajn (A/B test), ne regresiju.

Model je dobar koliko i podaci — vaš model ne može uhvatiti faktore koje niste mjerili poput kvalitete fotografije, trenutnih trendova, algoritamskih promjena, ili jednostavno sreće. Zato R-kvadrat nikad neće biti 1, i to je sasvim normalno.

Ekstrapolacija je opasna — model je treniran na podacima s 0 do 30 hashtagova. Što bi predvidio za 50 hashtagova? Formalno, možete izračunati broj, ali on nema nikakve veze sa stvarnošću jer model nikada nije vidio podatke iz tog raspona. Predviđanje izvan raspona vaših podataka je ekstrapolacija i trebate je izbjegavati.

Pretpostavke moraju biti zadovoljene — ako dijagnostički grafovi pokazuju ozbiljna odstupanja, nelinearnost, heteroskedastičnost, ili nenormalne reziduale, rezultati mogu biti nepouzdani. Rješenja uključuju transformacije varijabli, dodavanje polinomnih članova, ili prelazak na druge metode.

❗ Ključni zaključci

1. **Regresija modelira odnos** između prediktora (X) i ishoda (Y). Jednostavna: Y = b + b X. Višestruka: Y = b + b X + b X + …
2. `lm(y ~ x1 + x2, data)` provodi regresiju u R-u. `summary()` daje koeficijente, standardne pogreške, t-testove, p-vrijednosti, R^2 i F-test.
3. **Svaki koeficijent u višestrukoj regresiji je parcijalni efekt** - promjena Y za jediničnu promjenu X uz kontrolu svih ostalih prediktora. Ovo je ono "držeći sve ostalo jednakim."
4. **R-kvadrat** je udio varijabilnosti objašnjen modelom. Prilagođeni R^2 korigira za broj prediktora. U komunikologiji, R^2 između 0.10 i 0.30 je uobičajen.
5. **AIC** služi za usporedbu modela - niži AIC znači bolji model jer penalizira nepotrebnu kompleksnost.
6. **Četiri pretpostavke** (linearnost, nezavisnost, homoskedastičnost, normalnost reziduala) provjeravate dijagnostičkim grafovima pomoću funkcije `plot(model)`.
7. **VIF** mjeri multikolinearnost. VIF ispod 5 je prihvatljiv. VIF iznad 10 je problematičan i znači da su prediktori previše korelirani.
8. **Nelinearne odnose** možete uhvatiti polinomom: `I(x^2)` dodaje kvadratni član. LOESS krivulja otkriva nelinearnost vizualno.
9. **Standardizirani koeficijenti** (beta) stavljaju sve prediktore na istu skalu i omogućuju usporedbu relativne važnosti.
10. **Cookova udaljenost** identificira utjecajne točke. Prag: 4/n. Uvijek usporedite model s i bez utjecajnih točaka.
11. **Regresija nije kauzalnost** - otkriva samo asocijacije. Za kauzalne zaključke trebate eksperiment, a ekstrapolacija izvan raspona podataka je nepouzdana.
12. **Kompletni izvještaj** uključuje opis uzorka i modela, F-test i R^2, značajne koeficijente s interpretacijom, dijagnostiku pretpostavki, vizualizaciju efekata i praktične preporuke.

14 Zadaci za vježbu

1. Učitajte `social_engagement.csv`. Provedite jednostavnu regresiju `engagement_rate ~ num_hashtags`. Interpretirajte koeficijent i R^2. Pogledajte dijagnostičke grafove. Zatim dodajte kvadratni član `I(num_hashtags^2)` i usporedite dva modela po AIC-u i prilagođenom R-kvadratu.
2. Izradite višestruki model s barem 4 prediktora. Izračunajte VIF za numeričke prediktore.

Napišite rezultate u APA formatu koristeći obrazac iz poglavlja o izvještavanju.

3. Kreirajte graf koji prikazuje predviđeni engagement za svaku kombinaciju `content_type` i `topic` (pri prosječnim vrijednostima ostalih prediktora). Koja kombinacija je najuspješnija?

15 Dodatno čitanje

Obavezno

Navarro, D. (2018). *Learning Statistics with R*, Chapter 15 (Linear Regression). Besplatno dostupno na learningstatisticswithr.com. Pokriva jednostavnu i višestruku regresiju s R kodom i izvrsnim konceptualnim objašnjenjima.

Preporučeno

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd edition). Springer. Poglavlje 3. Besplatno na statlearning.com. Moderniji pristup regresijskom modeliranju s naglašenim vizualnim objašnjenjima.

Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd edition). SAGE. Referentni priručnik za regresijsku dijagnostiku u R-u. Korisno kad trebate detaljniju provjeru pretpostavki.

16 Pojmovnik

Pojam	Objašnjenje
Linearna regresija	Modeliranje linearne veze između prediktora (X) i ishoda (Y) kroz jednadžbu Y = b + b X + .
Jednostavna regresija	Regresija s jednim prediktorom.
Višestruka regresija	Regresija s dva ili više prediktora, gdje su koeficijenti parcijalni efekti.
Koeficijent (b, slope)	Promjena Y za jediničnu promjenu X uz kontrolu ostalih prediktora.
Intercept (b)	Predviđeni Y kad su svi prediktori jednaki nuli - često bez praktičnog značenja.
Parcijalni efekt	Efekt jednog prediktora uz kontrolu (držanje konstantnima) svih ostalih.
Rezidual (e)	Razlika između opaženog i predviđenog Y, što je označeno s e = Y − Ŷ.
R-kvadrat (R^2)	Udio varijabilnosti Y-a objašnjen modelom, gdje 0 znači da model ne objašnjava ništa, a 1 znači savršenost.

Pojam	Objašnjenje
Prilagođeni R^2	R^2 korigiran za broj prediktora, koristi se za usporedbu modela s različitim brojem prediktora.
AIC	Akaike Information Criterion, gdje niža vrijednost znači bolji model jer penalizira kompleksnost.
OLS	Ordinary Least Squares - metoda koja minimizira sumu kvadriranih reziduala.
Dummy varijabla	Binarna (0/1) varijabla za kodiranje kategorija, koju R automatski kreira u `lm()`.
VIF	Variance Inflation Factor, mjeri multikolinearnost gdje je VIF < 5 prihvatljivo, a > 10 problematično.
Multikolinearnost	Visoka korelacija između prediktora koja čini koeficijente nestabilnima.
Cookova udaljenost	Mjera utjecaja pojedinog opažanja na model, gdje je prag $4/n$.
Leverage	Mjera koliko je opažanje ekstremno u prostoru prediktora - visok leverage znači potencijalno utjecajno.
Standardizirani koeficijent (beta)	Koeficijent izražen u SD jedinicama što omogućuje usporedbu prediktora.
Polinomijalna regresija	Dodavanje kvadratnog (ili višeg) člana za uhvatiti nelinearne odnose pomoću `I(x^2)`.
Homoskedastičnost	Jednaka varijanca reziduala za sve predviđene vrijednosti, pretpostavka regresije.
Ekstrapolacija	Predviđanje izvan raspona podataka, nepouzdano jer model nije treniran za te vrijednosti.
LOESS	Locally Estimated Scatterplot Smoothing - fleksibilna krivulja za otkrivanje nelinearnih trendova.
`lm()`	R funkcija za linearnu regresiju, koristi se kao `lm(y ~ x1 + x2, data = ...)`.
`summary()`	Na `lm()` objektu daje koeficijente, SE, t, p, R^2 i F-test.
`predict()`	Generira predviđene vrijednosti pomoću `predict(model, newdata = ...)` za nova opažanja.
`residuals()`	Izvlači reziduale iz modela.
`broom::tidy()`	Pretvara model output u tibble s koeficijentima, SE, t, p i CI.

Pojam	Objašnjenje
`AIC()`	R funkcija za izračun AIC-a, koristi se kao `AIC(model1, model2)` za usporedbu.