Uvod u znanost o podacima

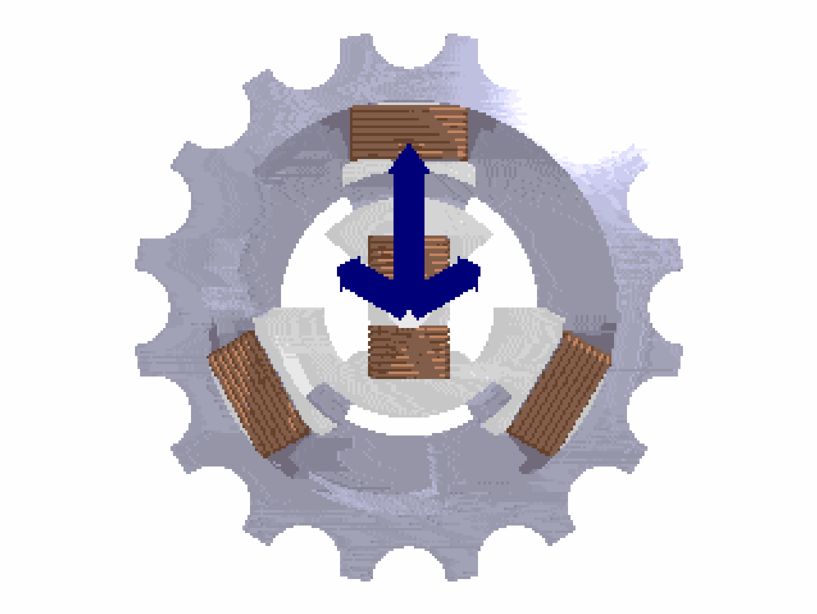
Uvod u regresijsku analizu

Bojana Dalbelo Bašić

5. Predavanje

ak. god. 2021./2022.

Univerzalni stroj, puno primjena, puno varijanti



Elektromotor - Mtodorov_69, CC BY-SA 3.0 via Wikimedia Commons

Linearna regresija

Generalizirani linearni modeli (**GLM**) (logistička regresija, Poissonova regresija)

Coxova regresija

Regularizirani modeli

Nelinearna regresija

Temelj za ovo predavanje

Ovo predavanje temelji se na knjizi A. Gelman and J. Hill, "*Data Analysis Using Regression and Multilevel/Hierarchical Models*", najviše na trećem i četvrtom poglavlju:

3. Linear regression: the basics, i
4. Linear regression: before and after fitting the model;

kao i na materijalima Roberta Westa, *Applied Data Analysis* (EPFL), Regression analysis.

Linearna regresija – poznato do sada

- **Dano:** n parova točaka (X_i, y_i), X_i je k-dimenzionalni vektor prediktora (značajki?/varijabli?), y_i je izlazna vrijednost, i-te točke
- **Cilj:** naći optimalne koeficijente $\beta = (\beta_1, ..., \beta_k)$ za aproksimaciju y *kao linearne funkcije vektor*:

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

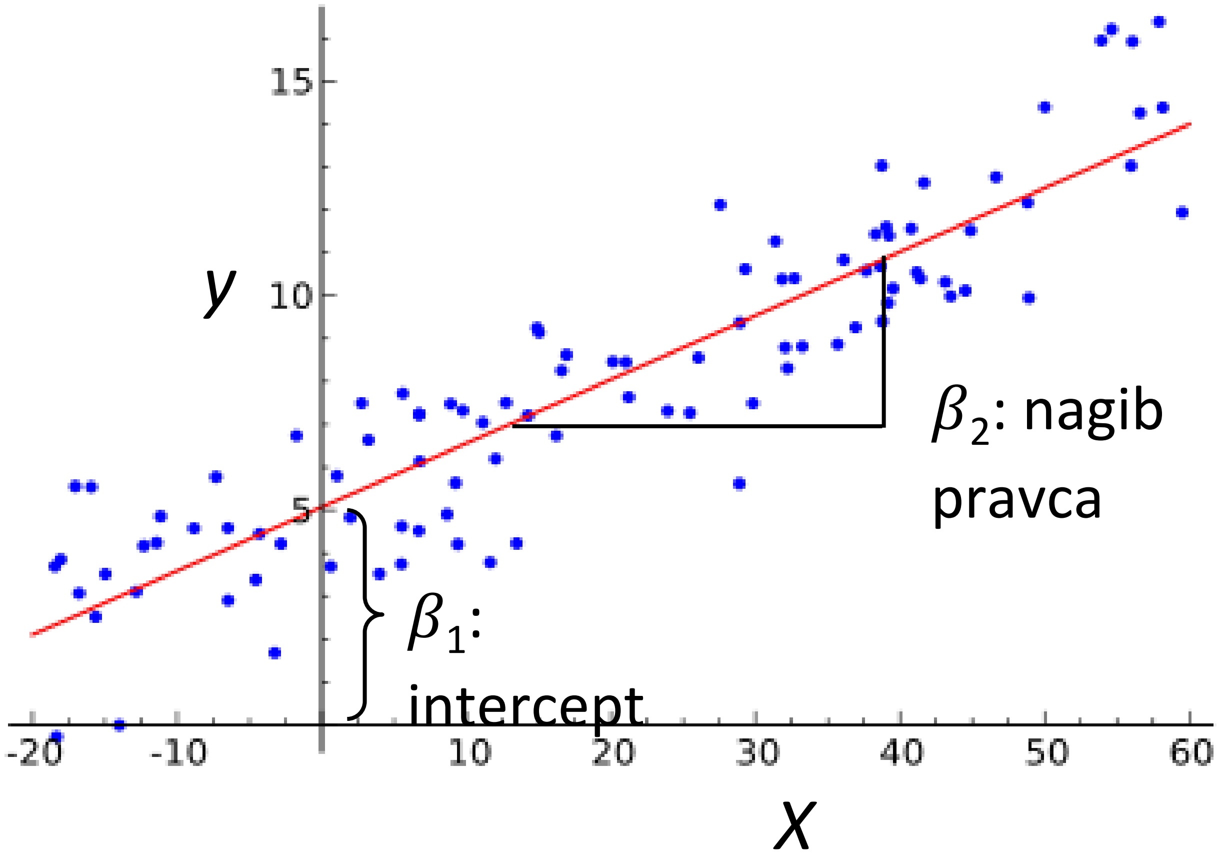
Skalarni product of dva vektora

gdje su ϵ_i pogreške (pretpostavke na ϵ_i ?)

- X_{i1} uobičajeno iznosi 1 $\Rightarrow$ β_1 je konstanta – intercept

Primjer: jedan prediktor

$$y \approx \beta_1 + \beta_2 X$$



Linearna regresija – poznato do sada

- **Dano:** n parova točaka (X_i, y_i), X_i je k-dimenzionalni vektor prediktora (a.k.a. značajki), y_i je izlazna vrijednost, i-te točke
- **Cilj:** naći optimalne koeficijente $\beta = (\beta_1, \ldots, \beta_k)$ za aproksimaciju *y kao linearne funkcije vektor*:

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

- X_{i1} uobičajeno iznosi 1 $\Rightarrow$ β_1 je konstanta – intercept

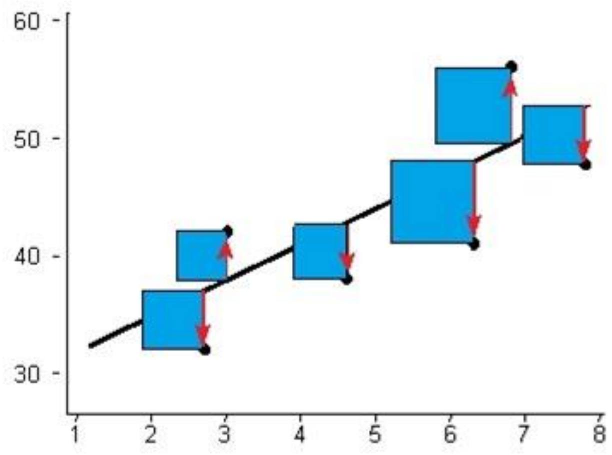
Kriterij optimalnosti: najmanji kvadrati

$$y_i \quad = \quad X_i\beta + \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

- Intuitivno, želimo da pogreške ϵ_i budu što manje
- Tehnički, želimo sumu kvadrata odstupanja što manju
  $\Leftrightarrow$ naći $\hat{\beta}$ tako da minimizira

$$\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$$

Rješenje $\quad \hat{\beta} \quad = \quad (X^TX)^{-1}X^TY$



Podsjetnik

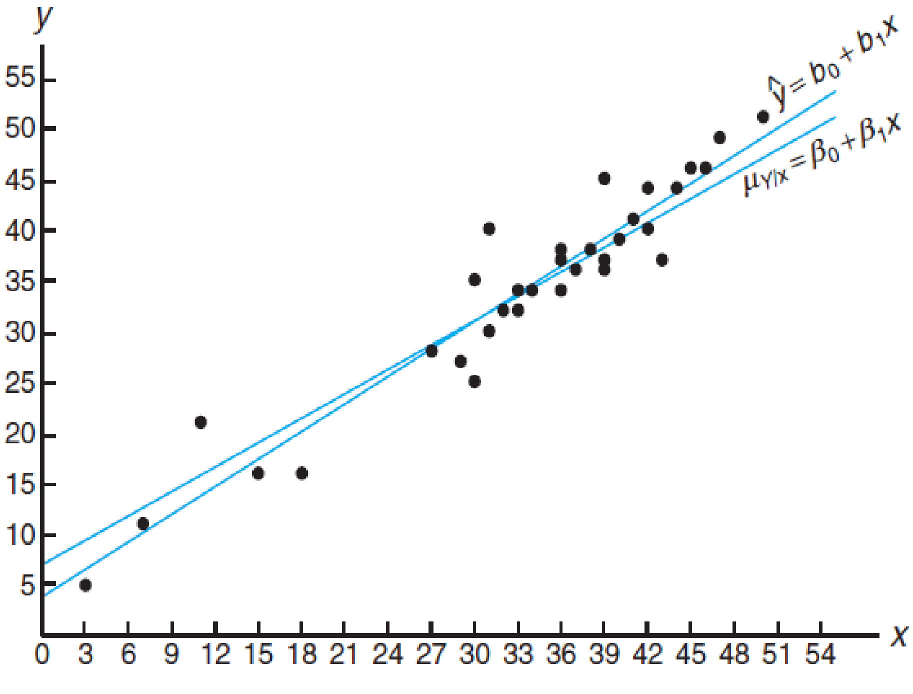


Figure 11.3: Scatter diagram with regression lines.

- Prava vrijednost parametara beta je nepoznata, kao i pogreška ϵ_i.

$\epsilon \sim N(0, \sigma^2)$, ϵ_i nezavisne

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Mi računamo procjene parametara β_i koje označavamo s β_i kapa ili b_i.

Podsjetnik
Razlika između reziduala e_i i pogreške ε_i

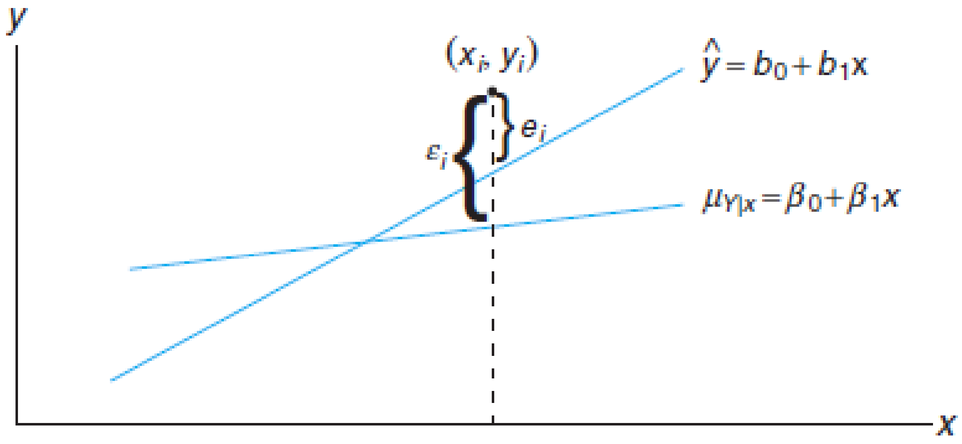


Figure 11.5: Comparing ϵ_i with the residual, e_i.

Za što koristimo regresiju?

- **Pedviđanje:** koristimo izračunati model da procijenimo izlaz y za novi X, koji do sada nije „viđen” u procesu izgradnje modela.

  Ako ste koristili regresiju do sada – to je bilo najvjerojatnije u kontekstu predviđanja

- **Deskriptivna analiza podataka:** usporedba srednjih vrijednosti kroz grupe podataka (DANAS!)
- **Modeliranje uzročnosti:** razumijevanje kako se izlaz y mijenja, ako manipuliramo prediktorima X. (ne nužno samo pomoću regresije, teme slijedećih poglavlja u knjizi)

Regresija kao usporedba srednjih vrijednosti izlaza

Primjer s jednim binarnim prediktorom X_i

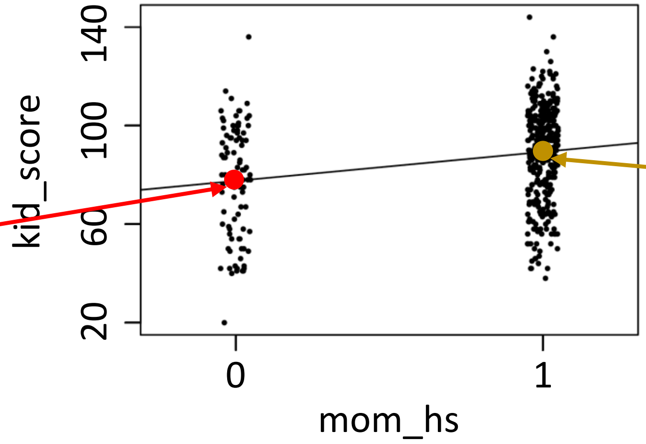
NE DA

- X_i = mom_hs = "Da li je mama završila fakultet?" $\in \{0, 1\}$
- y_i = kid_score = djetetov rezultat na kognitivnom testu $\in [0, 140]$

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

kid_score = 78 + 12 · mom_hs + error

Srednja vrijednost djetetovog rezultata za majke koje nisu završile fakultet: 78

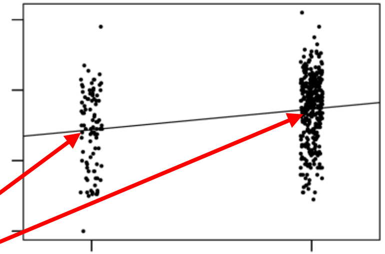


Srednja vrijednost djetetovog rezultata za majke koje jesu završile fakultet: 78 + 12 = 90

Jedan binarni prediktor X_i: Interpretacija procijenjenih prametara β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: srednja vrijednost za točke s $X_i = 0$
- **Nagib** (slope) β_2 : razlika u izlaznim vrijednostima između točaka s $X_i = 1$ i točaka s $X_i = 0$
- Objašnjenje: srednje vrijednosti minimiziraju kriterij najmanjih kvadrata



Zašto ne izračunati srednje vrijednosti odvojeno i usporediti ih

Primjer s jednim numeričkim kontinuiranim prediktorom X_i

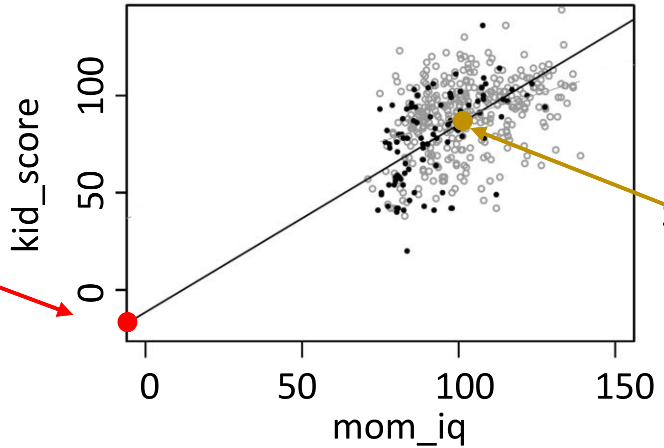
- X_i = mom_iq = majčin IQ rezultat $\in [70, 140]$
- y_i = kid_score = djetetov rezultat na kognitivnom testu$\in [0, 140]$

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

kid_score = 26 + 0.6 · mom_iq + error

(hipotetska) srednja vrijednost djetetovog rezultata *kid_score* za majke s IQ = 0:
26

?



Srednja vrijednost *kid_score* za majke s IQ = 100:
26 + 0.6 · 100 = 86

Jedan kontinuirani prediktor X_i: Interpretacija procijenjenih parametara β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

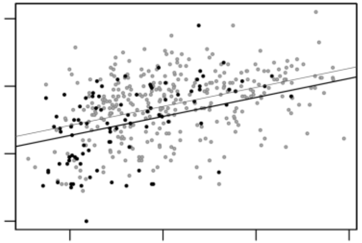
- **Intercept** β_1: prosječni izlaz za točke *i* with $X_i = 0$
- **Slope** β_2: razlika u izlazu između točaka čija se vrijednost X_i razlikuje za 1

Primjer s više prediktora

- (X_{i1} = 1 = constant)

  No Yes
- X_{i2} = mom_hs = “Da li je majka završila fakultet?” ∈ {0, 1}
- X_{i3} = mom_iq = majčin IQ rezultat ∈ [70, 140]
- y_i = kid_score = djetetov rezultat na kognitivnom testu ∈ [0, 140]

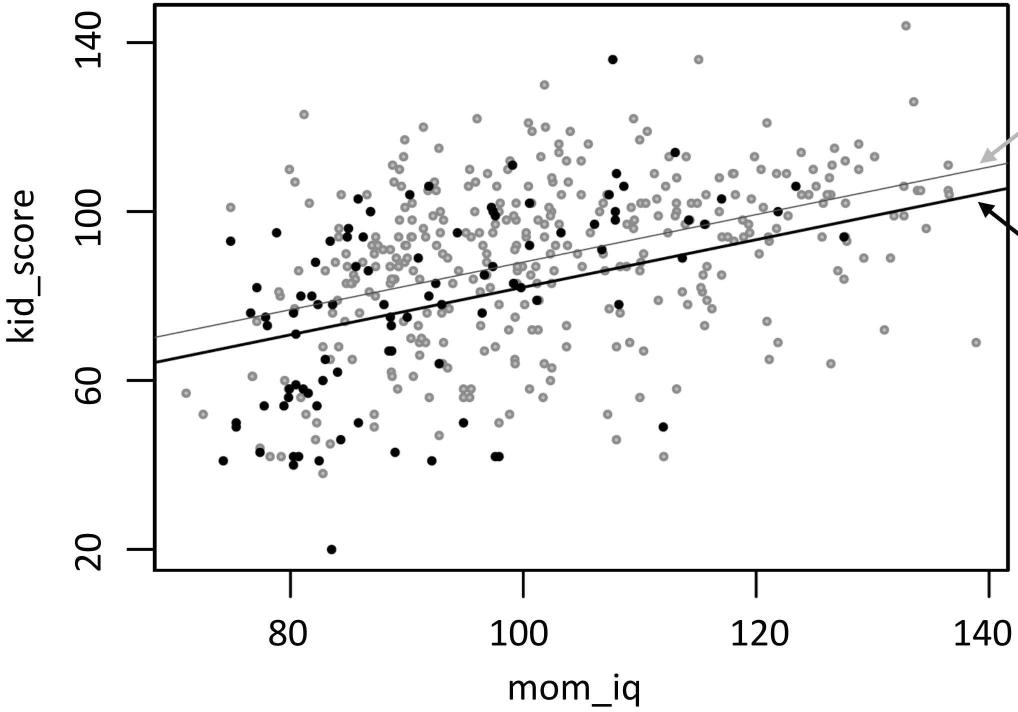
$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$



kid_score = 26 + 6 · mom_hs + 0.6 · mom_iq + error

Primjer s više prediktora

kid_score = 26 + 6 · mom_hs + 0.6 · mom_iq + error



Djeca čije majke jesu završile fakultet :
intercept = 26 + 6 = 32
nagib = 0.6

Djeca čije majke nisu završile fakultet :
intercept = 26
nagib = 0.6

Primjer s interakcijom prediktora

No Yes

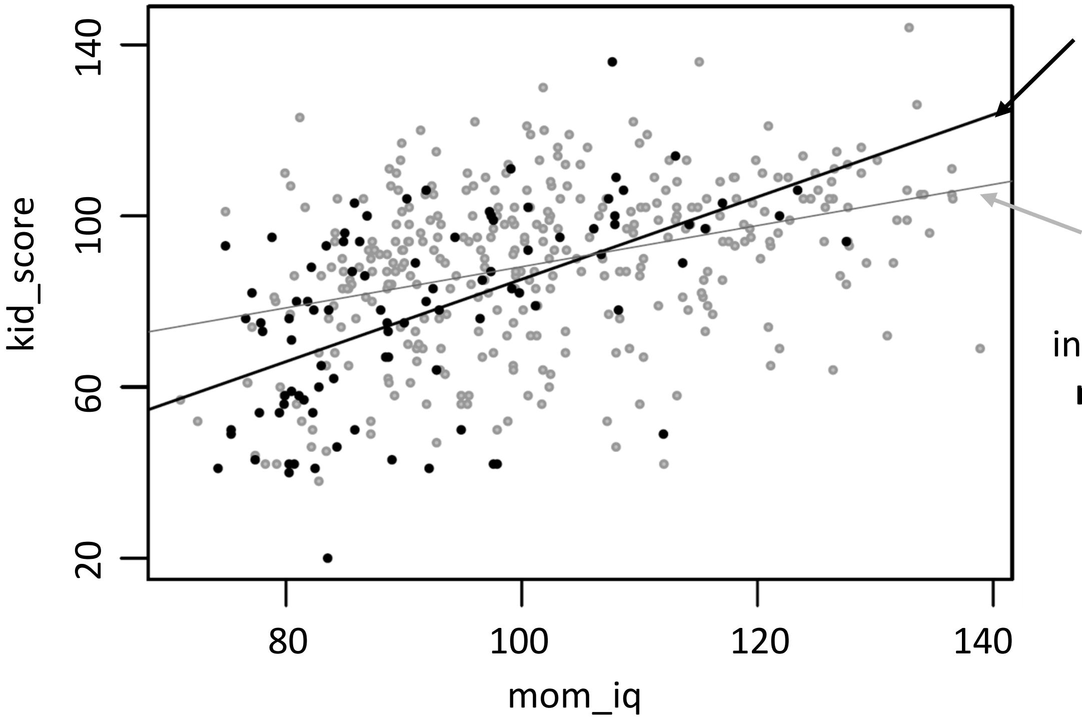
- X_{i2} = mom_hs = “Da li je majka završila fakultet?” $\in \{0, 1\}$
- X_{i3} = mom_iq = majčin IQ rezultat $\in [70, 140]$
- y_i = kid_score = djetetov rezultat na kognitivnom testu $\in [0, 140]$

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i2} X_{i3} + \epsilon_i$$

kid_score = −11 + 51 · mom_hs + 1.1 · mom_iq − 0.5 · mom_hs · mom_iq + error

Primjer s interakcijom prediktora

kid_score = −11 + 51 · mom_hs + 1.1 · mom_iq − 0.5 · mom_hs · mom_iq + error



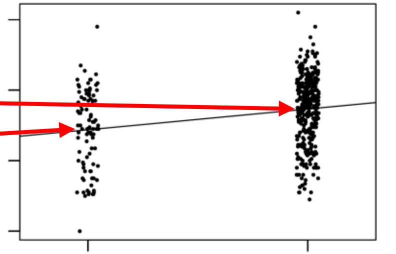
Djeca čije majke nisu završile fakultet :
intercept = −11
nagib = 1.1

Djeca čije majke jesu završile fakultet :
intercept = −11 + 51 = 40
nagib = 1.1 − 0.5 = 0.6

Zašto su nam interakcije važne?

- Mogućnost da model dobro opisuje različite podskupove koje imamo u podacima!
- U praksi: inputi koji imaju veliki učinak imaju tendenciju imati jake interakcije s drugim inputima (primjer: pušenje)
- Ipak ne mora biti isključivo tako…
- Modeli s interakcijom su lakše interpretabilni ako predprorcesiramo podatke (centriranje)

Zašto ne izračunati dvije srednje vrijednosti odvojeno i onda ih usporediti?



	Mame voze Mercedes	Mame ne voze mercedes
Mame su završile HS	avg kid_score 90	avg kid_score 90
Mame nisu završile HS	avg kid_score 78	avg kid_score 78

	Mame voze Mercedes	Mame ne voze mercedes
Mame su završile HS	990 žena	10 žena
Mame nisu završile HS	10 žena	990 žena

- Srednja vrijednost *kid_score* za Mercedes vozačice : $0.99 \cdot 90 + 0.01 \cdot 78 \approx 90$
- Srednja vrijednost *kid_score* za Mercedes ne-vozačice: $0.01 \cdot 90 + 0.99 \cdot 78 \approx 78$
- Ali vožnja Mercedesa uopće ne čini razliku (za fiksne HS prediktore)!
- Izvor zla: **korelacija** između završene HS i vožnje Mercedesa
- **Regresija kao spas : kid_score = 78 + 12 · mom_hs + 0 · mercedes + error**

	Mercedes	No Mercedes
Mame su završile HS	mean kid_score 90	mean kid_score 90
Mame nisu završile HS	mean kid_score 78	mean kid_score 78

	Mercedes	No Mercedes
Mame su završile HS	990 women	10 women
Mame nisu završile HS	10 women	990 women

Podsjetnik: Hi-kvadrat statistika

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

	mercedes	no mercedes	*Marginal Row Totals*
high school	990 (500) [480.2]	10 (500) [480.2]	1000
no high school	10 (500) [480.2]	990 (500) [480.2]	1000
Marginal Column Totals	1000	1000	2000 (Grand Total)

The chi-square statistic is 1920.8. The *p*-value is < 0.00001. Significant at *p* < .05.

Kvantificiranje neizvjesnosti

Kvantificiranje neizvjesnosti

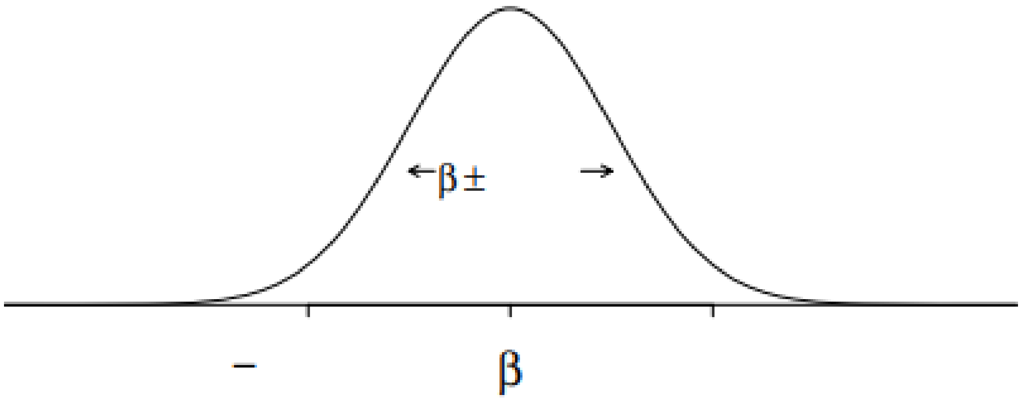
- Statistički software daje više od samih procjena koeficijenata β:

```
Residuals:
    Min      1Q  Median      3Q     Max
-52.873 -12.663   2.404  11.356  49.545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
mom.hs       5.95012    2.21181   2.690  0.00742 **
mom.iq       0.56391    0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-squared: 0.2141,     Adjusted R-squared: 0.2105
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

p-vrijednost: vjerojatnost procjene takvog koeficijenta ili ekstremnijeg ako je stvarni koeficijent nula (= H_0 hipoteza)

Pitanje

- *Uncertainty distribution* za koeficijente beta
- Koliko *Std. Error za …?*



Reziduali i R^2

- **Rezidual** za točku *i* : procjena pogreške i-te vrijednosti :

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

- Srednje vrijednost reziduala = 0

  (ukupna precijenjenost = ukupna podcijenjenost)

Reziduali i R^2

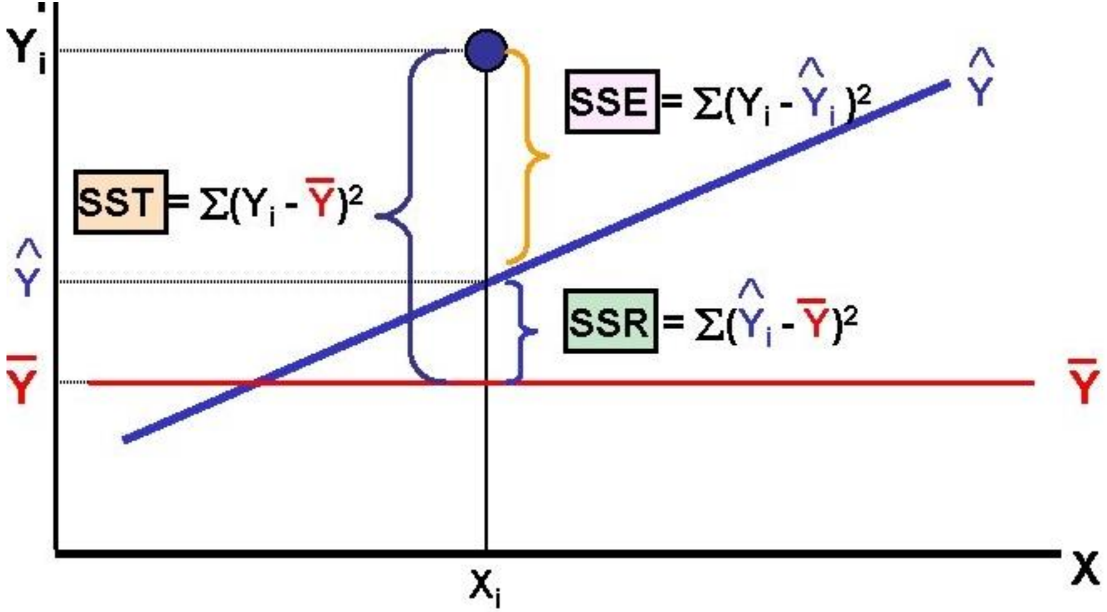
- **Rezidual** za točku *i* : procjena pogreške i-te vrijednosti :

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

- Srednje vrijednost reziduala = 0
(ukupna precijenjenost = ukupna podcjenjenost)
- Standardna devijacija reziduala
≈ procijenjena srednja vrijednost udaljenosti, predviđene vrijednosti od promatrane vrijednosti = "neobjašnjena varijabilnost"
- Udio varijance objašnjene modelom $R^2 = 1 - \hat{\sigma}^2 / s_y^2$

Varijanca izlaznih vrijednosti *y*

SST = SSR + SSE,

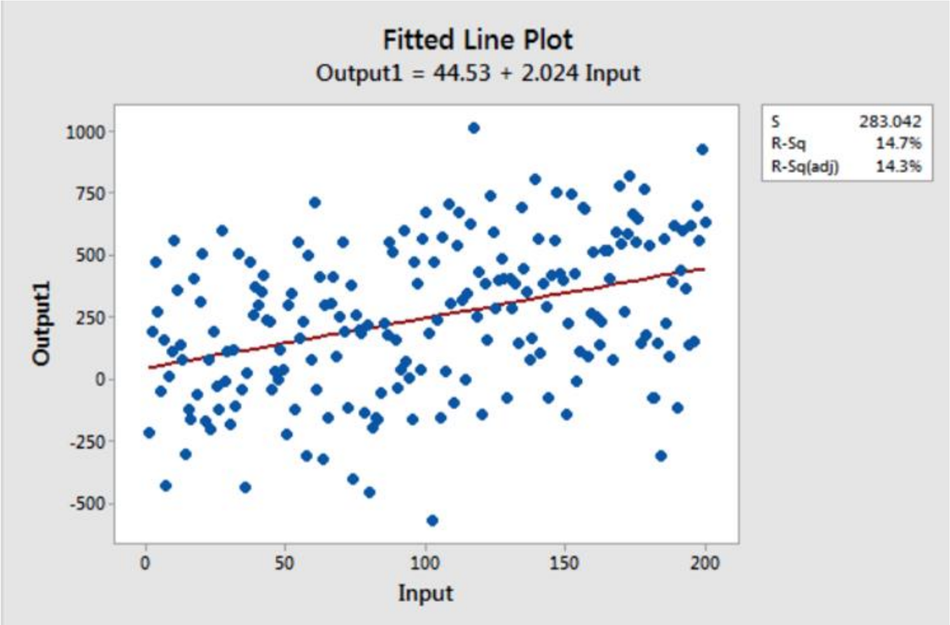


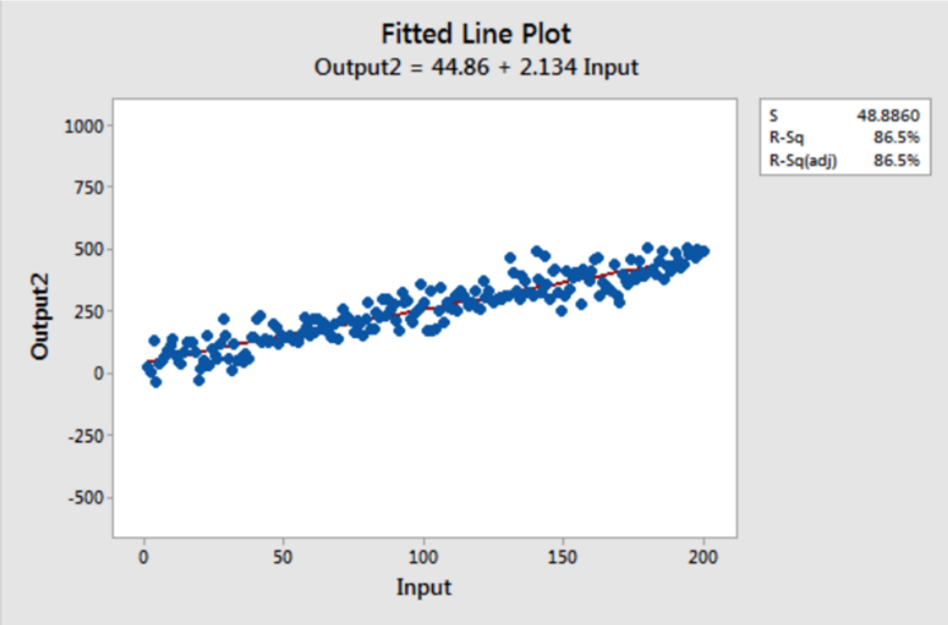
$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma\left(y - \bar{y}\right)^2}$$

R^2 = SSR/SST =1-SSE/SST, **$0 \leq R^2 \leq 1$**

Koeficijent determinacije: R^2

$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$

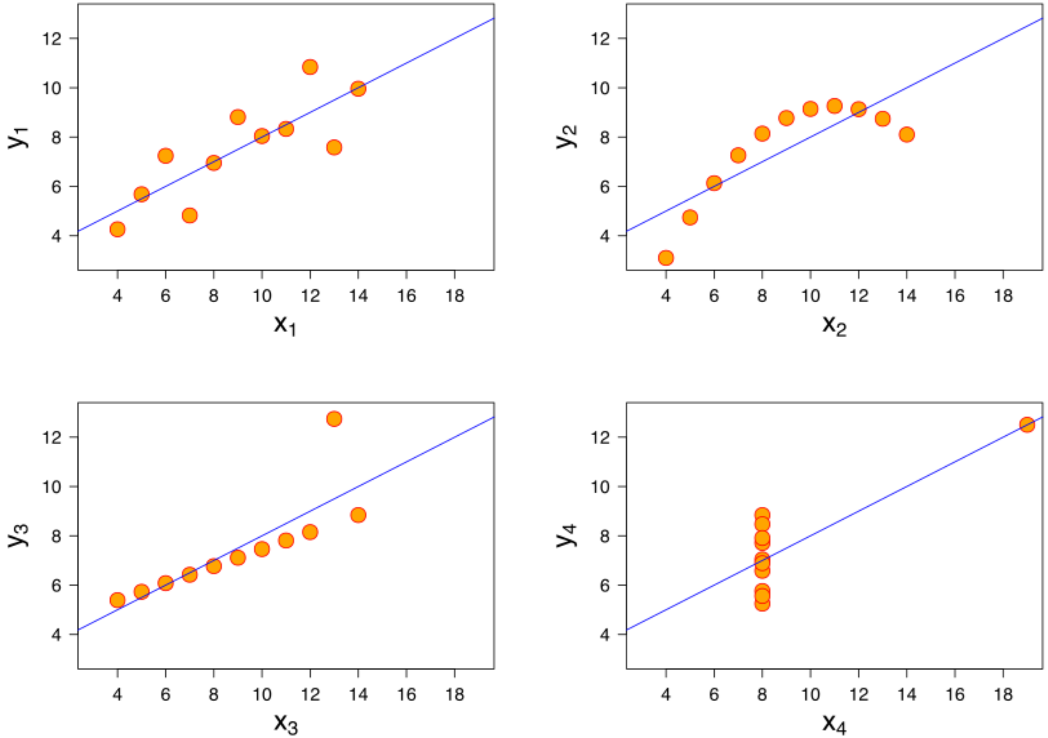




$R^2 = 0.147$

$R^2 = 0.865$

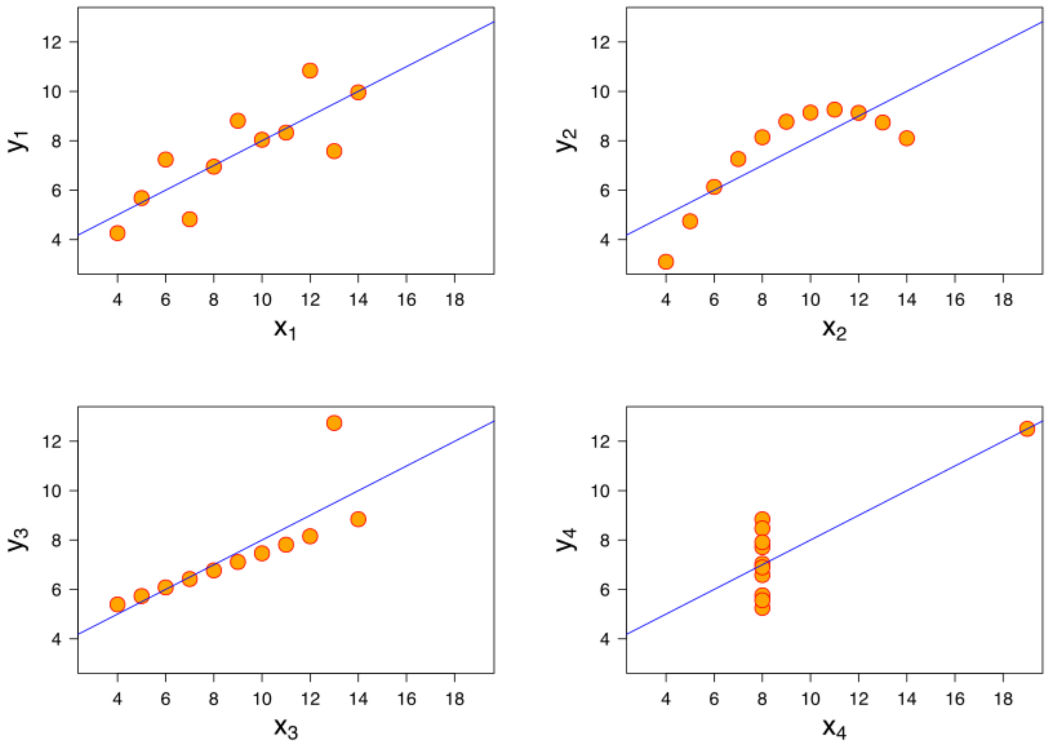
Koeficijent determinacije: R^2



Anscombe's quartet

Izvor: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

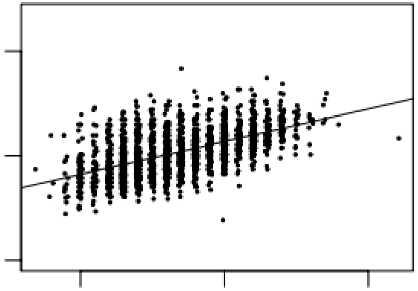
Koeficijent determinacije: R^2



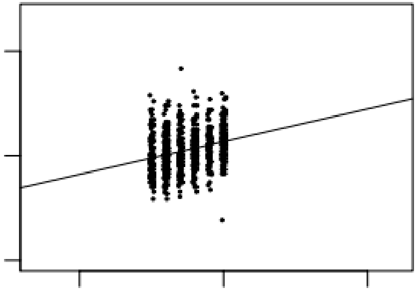
$R^2 = 0.67$ svugdje!

Koeficijent determinacije R^2

- Nije idealan: problem prenaučenosti – što više varijabli u modelu- bolji R^2 – prilagođeni R^2
- R^2 ne govori koliko je model blizu stvarnom!
- Primjer:



R^2 30%



R^2 10%

Pretpostavke u regresijskom modelu

Pretpostavke u regresijskog modela

1. Valjanost:
    a. Izlazne vrijednosti trebaju točno odražavati fenomen od interesa.
    b. Model treba uključivati sve relevantne prediktore
    c. Model treba generalizirati na slučajeve na koje će se primjenjivati

Pretpostavke regresijskog modela(2)

2. Aditivnost i linearnost:

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

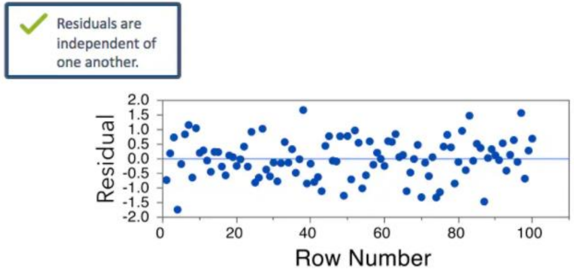
ali vrlo fleksibilna: linearan u prediktorima/koeficijentima (ne nužno u čistim ulaznim varijablama); prediktori mogu biti arbitrarne funkcije čistih ulaznih vrijednosti e.g.,

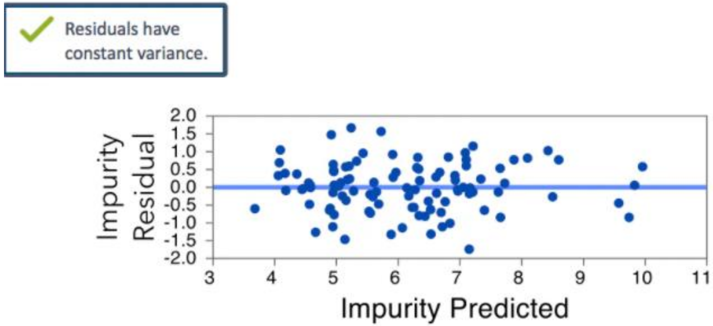
- log x, x^n, 1/x, …
- interakcije (i.e., produkti) višestrukih ulaza
- diskretizacija ulaza, kodiranog kao indikatorska varijabla

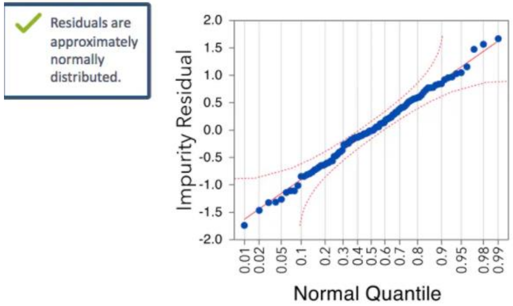
Pretpostavke regresijskog modela(3)

3. Nezavisnost pogrešaka: nema interakcije između ulaza
4. Konstantna varijanca reziduala
5. Normalnost reziduala

„Manje važno u praksi"







https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/mlr-residual-analysis-and-outliers.html

Transformacije prediktora i izlaza

Transformacije prediktora

- Kada primjenjujemo linearnu transformaciju na prediktore – model je i dalje linearan
- Procjene koeficijenata se mogu promijeniti, ali predviđanje izlaza i model ostaju nepromijenjeni.
- Primjer:

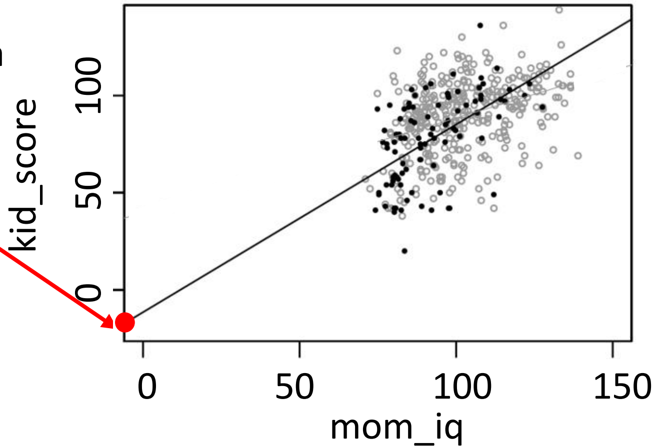
$$\begin{aligned}
\text{earnings} &= -61000 + 51 \cdot \text{height (in millimeters)} + \text{error} \\
\text{earnings} &= -61000 + 81000000 \cdot \text{height (in miles)} + \text{error.}
\end{aligned}$$

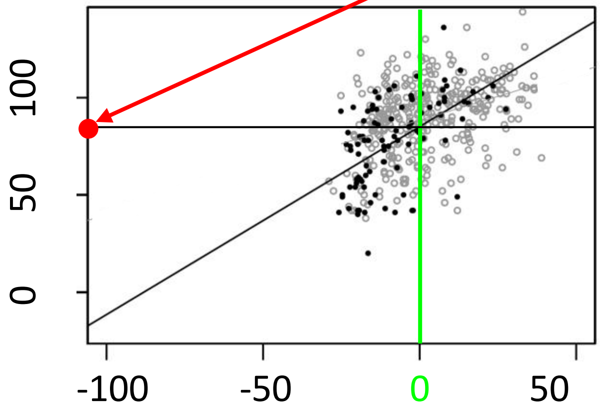
Prediktori centrirani oko srednjih vrijednosti

- Centrirati sve prediktore: izračunati srednju vrijednost i oduzeti je od svake vrijednosti prediktora: $X_{ik} \leftarrow X_{ik} - \text{mean}(X_{1k}, \ldots, X_{nk})$
- prediktor X_{ik} sada ima srednju vrijednost 0

Srednja vrijednost kid_score za mame sa IQ = 80

(pretpostavljena srednja vrijednost) kid_score za mame s IQ = 0: 26





Nakon centriranja prediktora oko srednje vrijednosti

… imamo pogodnu interpretaciju koeficijenata glavnih prediktora (glavni prediktori == non-interakcijski prediktori):

β_k = srednja vrijednost porasta izlaza y za svaku jedinicu porasta X_{ik}

kada svi drugi prediktori poprimaju svoju srednju vrijednost

Standardizacija via *z-scores*

- Prvo centrirati sve prediktore (oko srednje vrijednosti) i podijeliti ih sa svojom standardnom devijacijom
- $X_{ik} \leftarrow [X_{ik} - \text{mean}(X_{1k}, \ldots, X_{nk})] \,/\, \text{sd}(X_{1k}, \ldots, X_{nk})$

Standardizacija via *z-scores*

- Prvo centrirati sve prediktore (oko srednje vrijednosti) i podijeliti ih sa svojom standardnom devijacijom
- $X_{ik} \leftarrow [X_{ik} - \text{mean}(X_{1k}, \ldots, X_{nk})] / \text{sd}(X_{1k}, \ldots, X_{nk})$

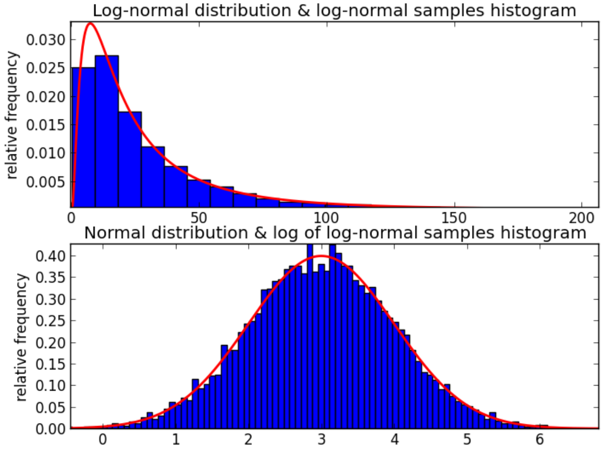
- Svi prediktori su u istim jedinicama ("**z-scores**"):
  udaljenost (izraženi u terminima standardne devijacije) od srednje vrijednosti.
- Omogućava nam usporedbu koeficijenata za prediktore sa prethodno neusporedivim jedinicama mjere, e.g., IQ score vs. zarada u eurima vs. visina u centimetrima

Logaritmi izlaznih vrijednosti

- **PRAKTIČNO:** ima smisla ako distribucija izlaznih vrijednosti ima „teške repove”

- Samo za ne-negativne izlaze

- **TEORIJSKI:** aditivni model postaje **multiplikativni:**



$$\log y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i$$

Exponentiating both sides yields

$$\begin{aligned} y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\ &= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i \end{aligned}$$

Logaritmi izlaznih vrijednosti: Interpretacija koeficijenata

$$
\begin{aligned}
y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\
&= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i
\end{aligned}
$$

- **Aditivno** povećanje od 1 u vrijednosti prediktora $X_{\cdot 1}$ povezano je s **multiplikativnim** povećanjem $B_1 = \exp(b_1)$ izlaznoj vrijednosti
- Ako $b_1 \approx 0$, odmah možemo interpretirati b_1 kao **relativno povećanje** u izlaznoj vrijednosti jer je $\exp(b_1) \approx 1 + b_1$
- Primjer: $b_1 = 0.05 \Rightarrow B_1 = \exp(b_1) \approx 1.05$
  $\Rightarrow$ “+1 in predictor $X_{\cdot 1}$” je pridruženo povećanju “+5% u izlazu”

Dalje od linearne regresije za usporedbu srednjih vrijednosti…

Što je dalje složenije od linearne regresije?

Generalizirani linearni modeli (GLM)

- Logistička regresija: binarni izlazi
- Poissonova regresija: ne-negativni cjelobrojni izlazi (e.g., brojevi)

Zaključak

- Linearna regresija može biti alat za usporedbu srednjih vrijednosti grupa
- Kako? Iščitati srednje vrijednosti grupa iz (*fitted*) koeficijenata.
- Prednosti pred čistom usporedbom srednjih vrijednosti “ručno”:
  - Uzima u obzir korelacije između prediktora
  - Kvantificiranje neizvjesnosti (značajnosti) “for free”
  - Aditivni ili multiplikativni modeli, treba uzeti log
- *Caveat emptor*:
  - Model mora biti adekvatno specificiran, inače besmisleni rezultati → budite kritični, napravite dijagnostiku (e.g., R^2)

Literatura

A. Gelman and J. Hill, “Data Analysis Using Regression and Multilevel/Hierarchical Models”

https://en.wikipedia.org/wiki/Difference_in_differences

Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye (2016.), *Probability and Statistics for Engineers and Scientists*