Vežbe 11

T-testovi parova

Nezavisni i upareni uzorci

- **Nezavisni** uzorci su dobijeni uzastopnim izvršavanjem eksperimenta (”merenja”) na **istom** objektu, uz pomoć **više** (”mernih”) instrumenata.
- **Upareni** uzorci su dobijeni uzastopnim izvršavanjem eksperimenta ("merenja") na **više** objekata, uz pomoć **jednog** (”mernog”) instrumenta.

- Primeri merenja dužine stranice kocke;
- Primeri ispitivanja uticaja leka (vs. placeba);
- Primer ocenjivanja TV programa u odnosu na pol.

- Za t-testove je bitno da su uzorci normalno raspodeljeni i da li im je varijansa jednaka ili ne. To se **najpre** ispituje!

T-test na nezavisnim uzorcima sa $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$ raspodelama

- Nulta hipoteza: $H_0(m_1 = m_2)$; Alternativna hipoteza: $H_1(m_1 \neq m_2)$.
- Input:

```
x1 ← c(...); x2 ← c(...);          #vektori uzorka
t.test(x1,x2, var.equal=...)       #na osnovu var.test-a
```

- Output:

```
        Welch Two Sample t-test
t = 16.069, df = 34.208, p-value < 2.2e-16
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
 12.18952 15.71815
 mean of x  mean of y
14.3869891  0.4331553
```

T-test na nezavisnim uzorcima sa $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$ raspodelama

- Nulta hipoteza: $H_0(m_1 = m_2)$; Alternativna hipoteza: $H_1(m_1 \neq m_2)$.
- Input:

```
Df ← data.frame(Vrednosti=x, Faktori=y))
t.test(Vrednosti~Faktori, data=Df, var.equal=...)
```

- Output:

```
        Welch Two Sample t-test
t = 16.069, df = 34.208, p-value < 2.2e-16
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
        12.18952 15.71815
        mean of x  mean of y
14.3869891  0.4331553
```

T-test na nezavisnim uzorcima sa $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$ raspodelama

- Nulta hipoteza: $H_0(m_1 = m_2)$; Alternativna hipoteza: $\mathbf{H_1(m_1 > m_2)}$.
- Input:

```
t.test(Vrednosti~Faktori, alt="greater", var.equal=...)
```

- Output:

```
        Two Sample t-test
t = -0.14095, df = 98, p-value = 0.5559
alternative hypothesis: true difference in means is
greater than 0; 95 percent confidence interval:
 -0.387991        Inf
sample estimates:
mean in group a mean in group b
     0.00306111      0.03341797
```

T-test na uparenim uzorcima sa $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$ raspodelama

- Nulta hipoteza: $H_0(m_1 = m_2)$; Alternativna hipoteza: $H_1(m_1 \neq m_2)$.
- Input:

```
D ← data.frame(Vrednosti=x.df, Faktori=y.df))
t.test(Vrednosti~Faktori, data=D, var.equal=..., paired=T)
```

- Output:

```
        Paired t-test
t = -0.51001, df = 49, p-value = 0.6123
alternative hypothesis: true difference in means is
not equal to 0; 95 percent confidence interval:
 -0.4578438  0.2724932
sample estimates:
mean of the differences
            -0.09267534
```

Two-sample permutation test

Two-sample permutation test

- Postoji objektivna, numerička razlika u uzoračkim sredinama, ali da li je sama ta vrednost razlike slučajna? Ako nije, sigurno postoji prava razlika.
- Primer razlike prosečne visina 15 momaka i 15 devojaka. Ako zaista ne postoji razlika, tih 30 brojeva kao da je dobijeno iz jedne populacije u kojoj su na slučajan način petnaestoro označeni kao "M" i petnaestoro kao "F".

- Glavna ideja: Iz populacije od 30 ljudi izabrati na slučajan način petnaestoro i labelovati ih kao "M", a ostale kao "F". Potom utvrditi koliko iznosi $\bar{x}_M - \bar{x}_F$ i da li je ta vrednost veća od originalne vrednosti razlike. Ovo uraditi "puno" puta.
- Nulta hipoteza je tvrdnja da ne postoji razlika u visinama (razlika je slučajna), a alternativna (one-sided ili two-sided) da razlika nije slučajna.

Two-sample permutation test

- Ako je alternativna hipoteza jednostrana, $H_1(m_1 - m_2 > 0)$, onda se p-vrednost računa kao udeo razlika većih ili jednakih realizovanoj vrednosti test statistike ("originalne" razlike).
- Ako je alternativna hipoteza jednostrana, $H_1(m_1 - m_2 < 0)$, onda...videti domaći.

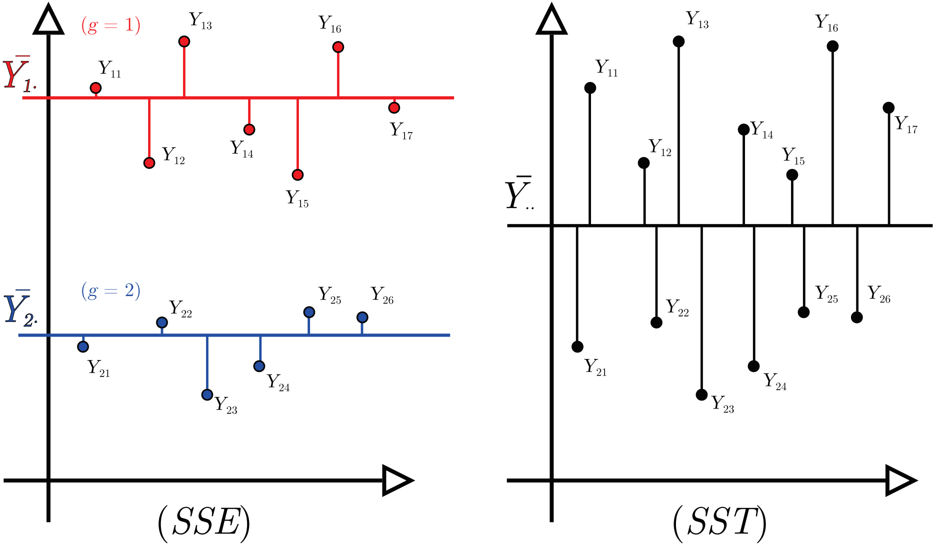
- Ako je alternativna hipoteza dvostrana, $H_1(m_1 - m_2 \neq 0)$, tj. $H_1(m_1 \neq m_2)$, onda se minimum od obe "jednostrane" p-vrednosti množi sa 2 (i eventualno zaokružuje na 1, ako je ta vrednost veća od 1). Tako dobijena vrednost predstavlja traženu p-vrednost ovog testa.
- Ako ne postoji jasan razlog zašto bi se koristio jednostrani test, uvek se koristi dvostrani test.

Two-sample permutation test, primer

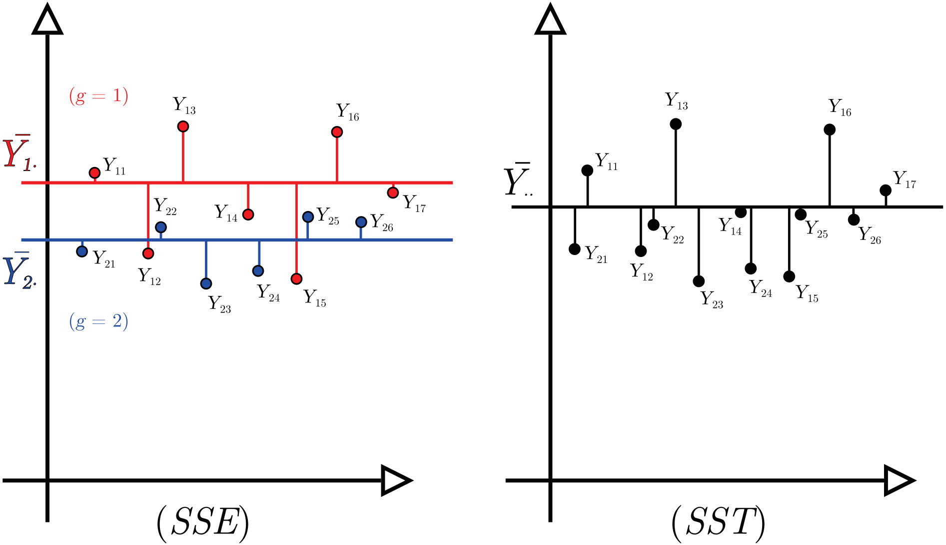
```
x.M ← c(178, 167, 172, ..., 176, 190, 169, 176, 180, 185)
x.F ← c(171, 159, 162, ..., 166, 162, 155, 158, 165, 170)
originalna.razlika ← mean(x.M)-mean(x.F)

x ← c(x.M, x.F)                        #sve visine
N ← 9999                               #proizvoljno
razlike ← numeric(N)                   #inicijalizacija
for (i in 1:N) {                       #prvih N razlika
        indeksi ← sample(30, size=15, replace=F)
        razlike[i] ← mean(x[indeksi])-mean(x[-indeksi])
}
razlike[N+1] ← originalna.razlika      #finalna razlika
(sum(razlike>=originalna.razlika))/(N+1)     #p-vrednost
```

Analiza varijanse - ANOVA

Analiza Varijanse, one-way ANOVA - ideja



Analiza Varijanse, one-way ANOVA - ideja



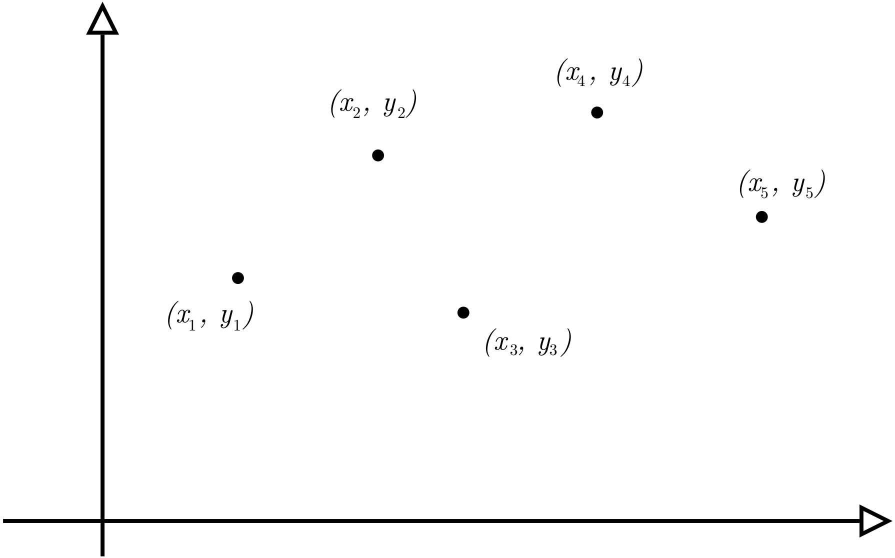
Analiza Varijanse, one-way ANOVA

- ANOVA se koristi samo ako su uzorci realizovani iz normalne raspodele!

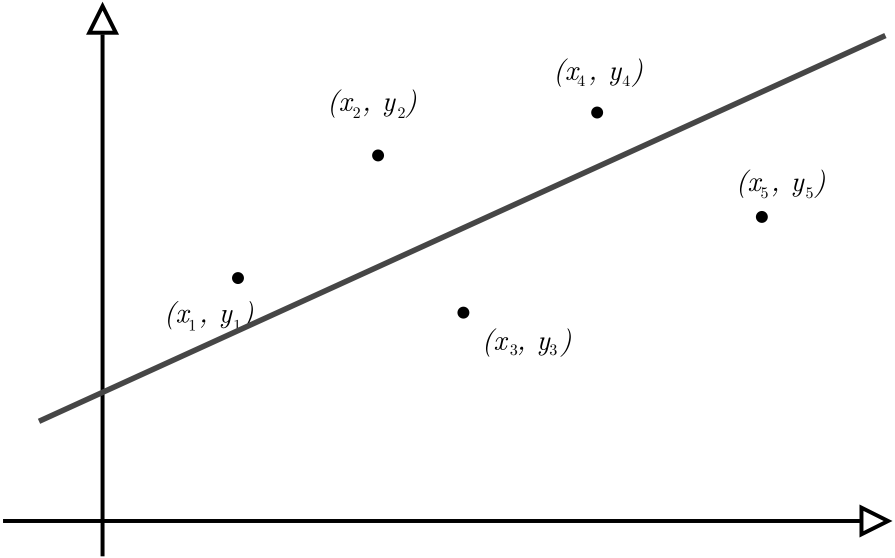
```
#generalizacija t.test-a:
oneway.test(Brojevi~Slova, data=Df)        #nejednake varijanse

#klasična oneway ANOVA:
oneway.test(Brojevi~Slova, data=Df, var.equal=TRUE)

m ← aov(Brojevi~Slova, data=Df)            #jednake varijanse
summary(m)                                  #tabela ANOVE

#neparametarski Kruskal-Wallisov test:
kruskal.test(Brojevi~Slova, data=Df)       #medijane
```

Regresija

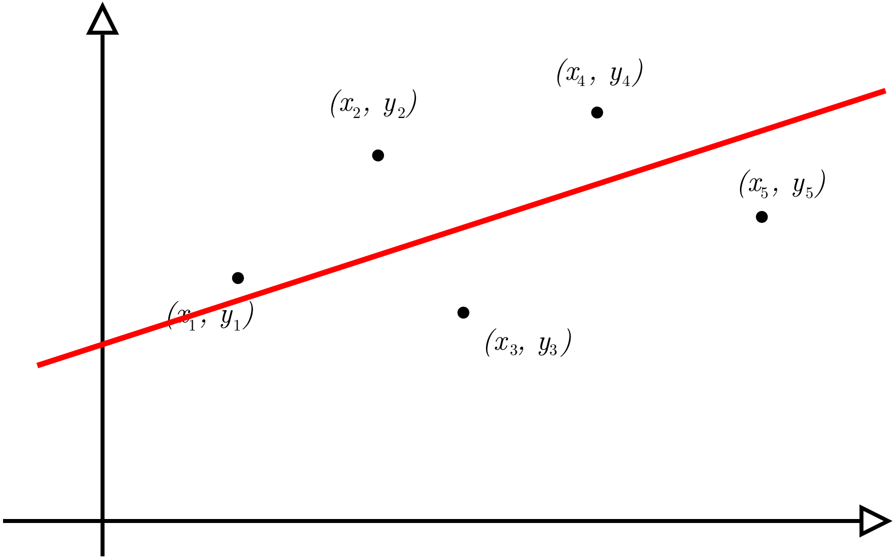
Regresija najmanjih kvadrata



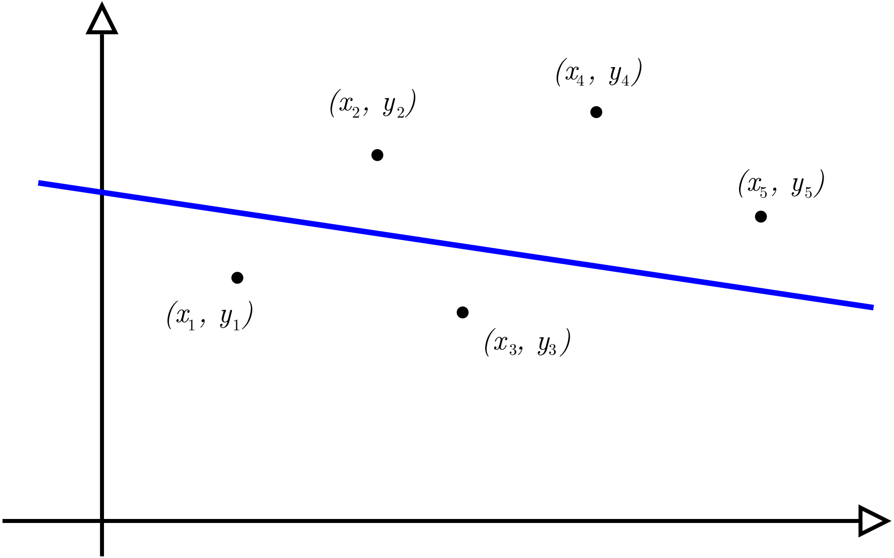
Regresija najmanjih kvadrata – ova prava?



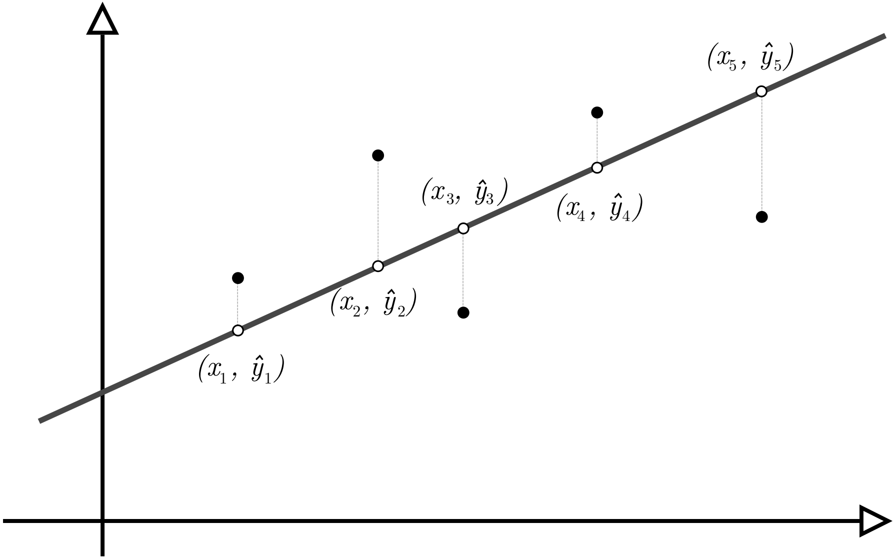
Regresija najmanjih kvadrata – ili ova?



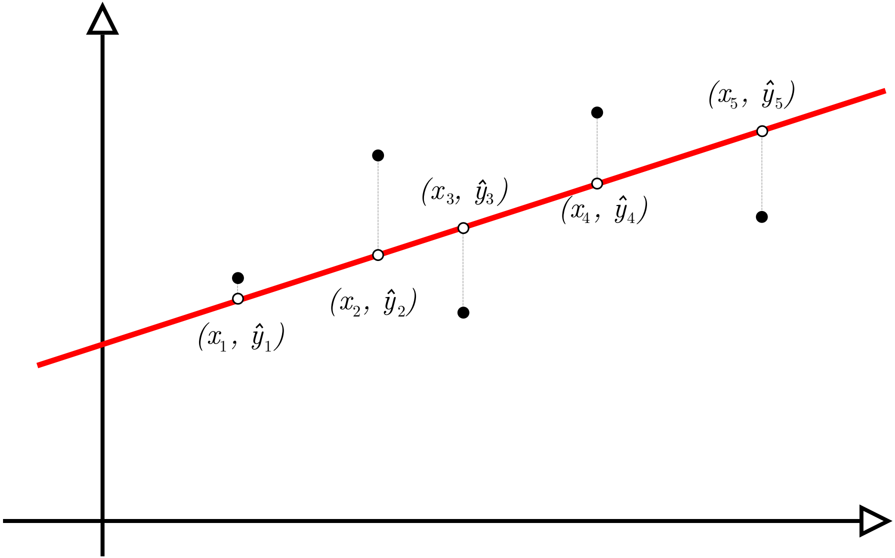
Regresija najmanjih kvadrata – ili ova?



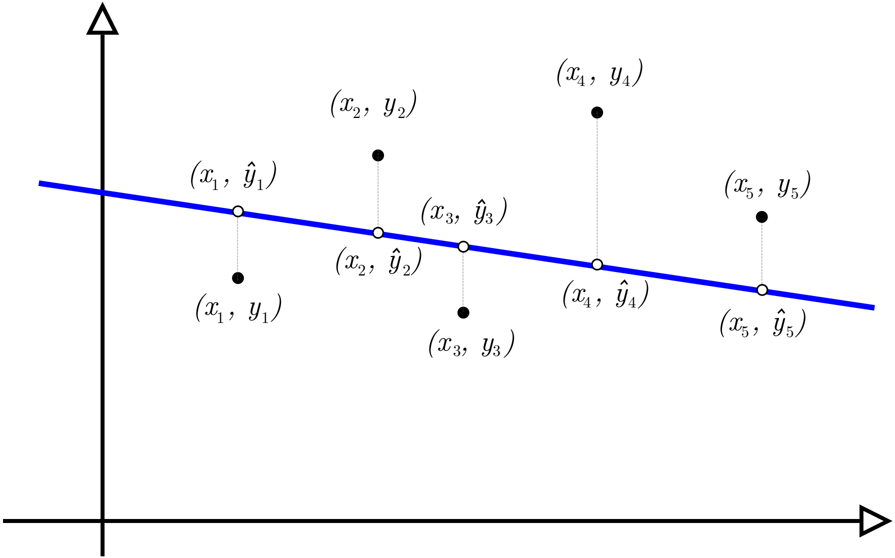
Regresija najmanjih kvadrata



Regresija najmanjih kvadrata



Regresija najmanjih kvadrata



Regresija najmanjih kvadrata

Za parove tačaka $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, regresiona prava $y = a + bx$ je ona koja minimizuje sumu kvadrata odstupanja vrednosti oblika $(y_i - \hat{y}_i)^2$:

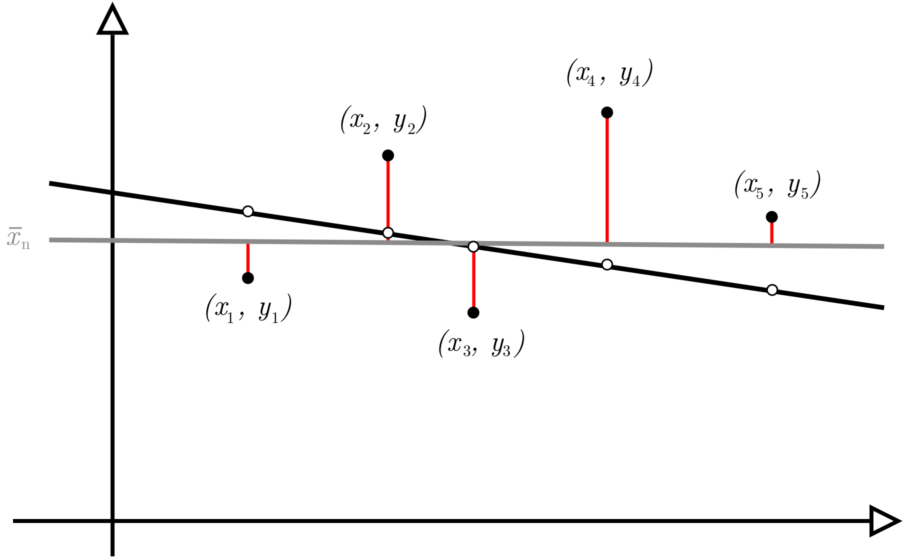
$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - (a + bx_i))^2 = (y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdot + (y_n - a - bx_n)^2,$$

posmatranu kao funkciju dve promenjive (a i b). Izjednačavanjem prvih parcijalnih izvoda sa nulom dobija se da je:

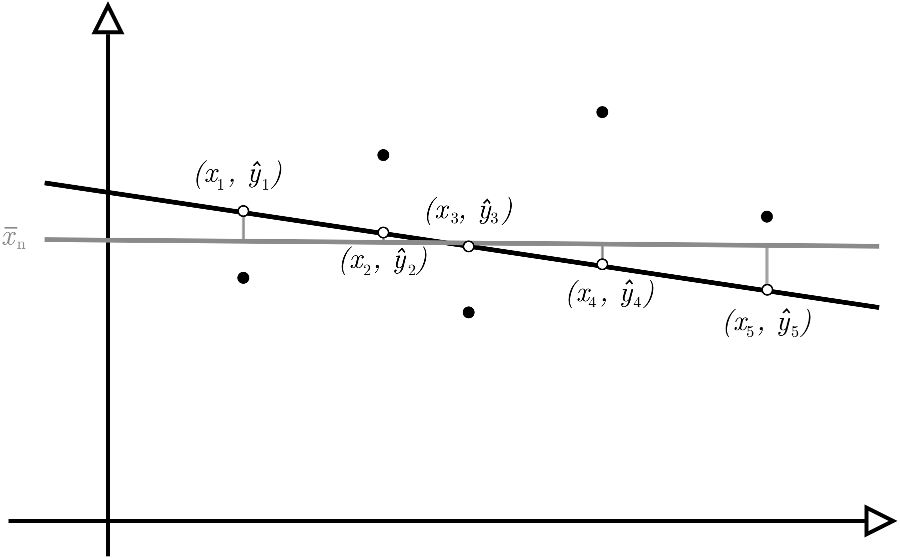
$$b = \frac{\sum (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum (x_i - \bar{x}_n)^2} = \frac{s_{xy}}{ss_x},$$

a, na osnovu činjenice da tačka $(\bar{x}_n, \bar{y}_n)$ pripada regresionoj pravoj, sledi i da je $\bar{y}_n = a + b\bar{x}_n$, tj. $a = \bar{y}_n - b\bar{x}_n$.

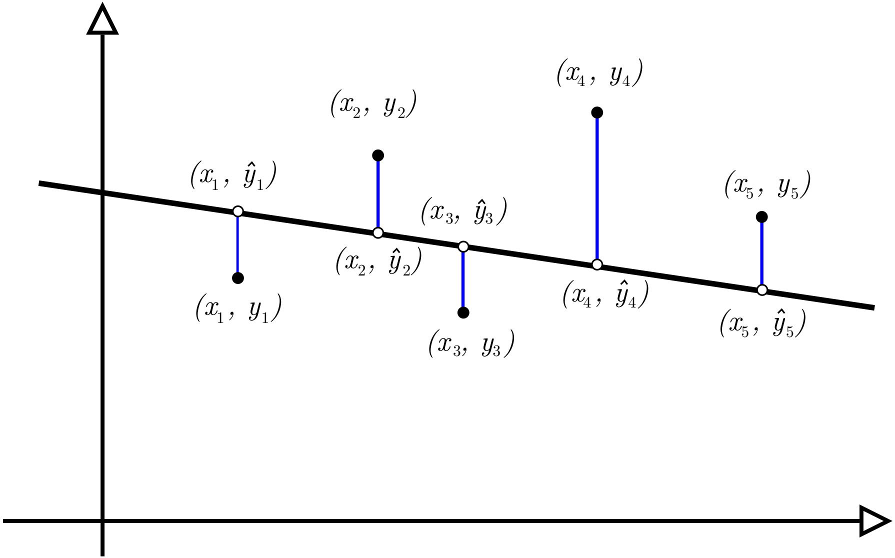
Regresija najmanjih kvadrata



Regresija najmanjih kvadrata



Regresija najmanjih kvadrata



lm(·)

```
lm(y~x)                         #linear model
lm(y~x, data=TX)                #TX je ime dataframe-a
lm(y~x+0, data=TX)              #model y = β1 * x + ε
lm(y~u+v+w, data=TX)            #u, v, w su sve prediktori

m ← lm(y~u+v+w)
coef(m)                         #koeficijenti regresije
confint(m)                      #int. poverenja za koef. reg.
deviance(m)                     #suma kvadrata reziduala
fitted(m)                       #fitovane vrednosti
residuals(m)                    #reziduali
summary(m)
```

Pretpostavke jednostavnog linearnog modela (JLM)

Ako posmatramo $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, gde su x_i realni brojevi, a Y_i slučajne promenjive koje predstavljaju predikciju $\hat{y}_i$ za x_i, onda su pretpostavke jednostavnog linearnog modela:

- vrednosti x_i su fiksirane;
- veza između očekivane vrednosti Y_i i x_i je linearna, $\mu_i = E(Y_i|x_i) = \alpha + \beta x_i$;
- reziduali $\varepsilon_i = Y_i - \mu_i$ su nezavisni;
- reziduali imaju konstantu (i jednaku) varijansu σ^2 ("homoskedastičnost");
- reziduali su saglasni sa normalnom raspodelom.

Pod ovim pretpostavkama su ocenjivači $\hat{\alpha}$, $\hat{\beta}$ i $\hat{\sigma}^2$ (dobijeni MMV) dati sa:

$$\hat{\beta} = \frac{\sum(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum(x_i - \bar{x}_n)^2}, \qquad \text{i} \qquad \hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n, \qquad \text{i} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2,$$

gde su $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$.

Jednostavan linearni model, JLM

Pod pretpostavkama JLM, Y_i su nezavisne slučajne promenljive saglasne sa normalnom $\mathcal{N}(\mu_i, \sigma)$. Takođe, ocenjivači $\hat{\beta}$ i $\hat{\alpha}$ su nezavisne slučajne promenjive i čak važi:

$$\hat{\beta} : \mathcal{N}\left(\beta, \frac{\sigma}{\sqrt{ss_x}}\right), \qquad \text{i} \qquad \hat{\alpha} : \mathcal{N}\left(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}_n^2}{ss_x}}\right), \qquad \text{i} \qquad n\hat{\sigma}^2/\sigma^2 : \chi^2_{n-2}.$$

Uvodeći oznaku $S^2 = \frac{n}{n-2}\sigma^2$, i primetivši da je:

$$\frac{\hat{\beta} - \beta}{\sigma/\sqrt{ss_x}} : \mathcal{N}(0, 1), \qquad \text{i} \qquad n\hat{\sigma}^2/\sigma^2 = (n-2)S^2/\sigma^2 : \chi^2_{n-2},$$

sledi da količnik:

$$\frac{\frac{\hat{\beta}-\beta}{\sigma/\sqrt{ss_x}}}{\sqrt{\frac{n\hat{\sigma}^2/\sigma^2}{n-2}}} = \frac{\frac{\hat{\beta}-\beta}{\sigma/\sqrt{ss_x}}}{\sqrt{\frac{(n-2)S^2/\sigma^2}{n-2}}} = \frac{\hat{\beta} - \beta}{S/\sqrt{ss_x}} : t_{n-2}.$$

Ovo je test statistika kojom se testira hipoteza $H_0(\beta = 0)$. vidi output komande `summary(lm(·))`.

Modeli

- Modeli:
  - daju vezu između varijabli;
  - omogućavaju predikcije.
- Prosta linearna regresija predstavlja regresiju oblika:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i \in \{1, 2, \ldots, n\},$$

pri čemu su poznati $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$. Osnovni zadatak je odrediti koeficijente u modelu ("fitovati").

- Pitanja:
  - da li je model statistički značajan?
  - da li su (svi) koeficijenti statistički značajni?
  - da li je model koristan?
  - da li model dobro fituje podatke?
  - da li podaci zadovoljavaju pretpostavke linearne regresije?

Modeli - II

- Da li je model statistički značajan?
  - F-statistika u `summary`-ju.
- Da li su (svi) koeficijenti statistički značajni?
  - t-statistike i p-vrednosti.
- Da li je model koristan?
  - koeficijent determinacije R^2.
- Da li model dobro fituje podatke?
  - plot reziduala.
- Da li podaci zadovoljavaju pretpostavke linearne regresije?
  - plot i test outlajera.

summary(lm(·)), primer proste regresije

```
Residuals:
    Min      1Q  Median      3Q     Max
 -8.390  -4.374  -3.162  -0.285 154.604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.320      1.625   6.966 3.77e-10 ***
          x   17.464      1.623  10.762  < 2e-16 ***

Residual standard error: 16.22 on 98 degrees of freedom
Multiple R-squared:  0.5417,    Adjusted R-squared:  0.537
F-statistic: 115.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

summary(lm(·)), primer multiple regresije

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.3965  -0.9472  -0.4708  1.3730  3.1283
Coefficients:
            Estimate Std.  Error t value Pr(>|t|)
(Intercept)   1.4222 1.4036   1.013  0.32029
          u   1.0359 0.2811   3.685  0.00106 **
          v   0.9217 0.3787   2.434  0.02211 *
          w   0.7261 0.3652   1.988  0.05744 .

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981, Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF, p-value: 0.0003915
```

Reziduali

- U idealnom slučaju:
  - Residuals vs. Fitted: ne pokazuje nikakvu očiglednu zavisnost;
  - Reziduali slede normalnu raspodelu (Q-Q plot);
  - Scale-Location: tačke su grupisane uniformno blizu “horizontalne”crvene linije.

```r
m ← lm(y~x)

plot(m, which=1)     #plot reziduala, Residuals vs. Fitted
plot(m, which=3)     #Scale-Location plot

library(car)
outlier.test(m)      #identifikacija potencijalnih outliera

library(lmtest)
dwtest(m)            #Durbin-Watsonov test autokorelacije
```

Predikcije, primer

- Najpre se formira novi dataframe, i "kolone" mu se nazovu kao vektori (prediktora) u modelu;

```
y ← sort(rnorm(35))                         #proizvoljni vektori
u ← sort(rchisq(35, 15))                    #koji se koriste
v ← sort(rpois(35, 4))                      #u ovom primeru
w ← sort(runif(35, 0, 1))                   #regresije

m ← lm(y~u+v+w)
Vred ← data.frame(u=10, v=4, w=0.5)         #kreiranje dataframea

predict(m, Vred)                            #predikcija za Vred
#95%-ni interval poverenja za predikciju u=10, v=4, w=0.5:
predict(m, Vred, interval="confidence", level=0.95)
```

Domaći rad

1. Googleovati šta je to `two-way` ANOVA i kako se koristi u R-u.
2. Kako bi se računala p-vrednost kod permutacionog testa, ako je alternativna hipoteza $H_1(m_1 - m_2 > 0)$?(*hint:* Chiara)
3. Kako se menja p-vrednost ako se broj uzorkovanja (`N`) povećava kod permutacionog testa? Zašto?
4. Istražiti kako se permutacioni test može koristiti za ispitivanje razlike u medijanama dva uzorka.
5. Navesti dva načina na koji se može izvesti kvadratna regresija.
6. Ispitati kojim argumentom bi se u t-testu mogla ispitati hipoteza da je npr. $m_1 - m_2 = 2$.
7. U t-testu, koje su sve vrednosti moguće za parametar "alt"?
8. Na slajdu 15 (primer permutacionog testa), kako glasi nulta hipoteza?
9. Navesti dve sličnosti i dve razlike između ANOVA testa i t-testa.

(budite
vredni)