Regresia Liniara Multipla

In acest document ne ocupam de regresia liniara multipla. Vom ilustra problematica cu exemplul utilizat la curs privind variabila dependenta=Inaltimea studentului/elevului in relatie de dependenta fata de variabile explicative (regresori) precum X1=Sexul, X2=Inaltimea Mamei si X3=Inaltimea Tatalui.

Pentru ca nu avem la dispozitie un set de date reale va trebui sa creeam noi variabilele. Cu R asta se poate face asa:

```
N<-1000;vecbeta<-c(0.5,-0.9,1.5)
```

Asadar am ales o populatie de talie 1000 si vectorul $\boldsymbol{\beta} = (0.5, -0.9, 1.5)^t$. Apoi generez valorile celor trei regresori cu functiile de generare aleatorie ale lui R.

```
Sex<-rbinom(N,size=1,prob=0.2)
Tata<-rnorm(N,mean=170,sd=sqrt(20))
Mama<-rnorm(N,mean=165,sd=sqrt(10))
```

Pentru Sex (variabila binara 0/1) am folosit functia de generare a distributiei Bernoulli. Am ales o probabilitate egala cu 0.2 ceea ce inseamna ca in populatie procentul de baieti va fi egal cu 20%. Pentru inaltimile tatilor si mamelor am folosit distributii normale de medii 170 respectiv 165 si de variante 20 respectiv 10. Inaltimile le consider ca fiind masurate in cm. Construiesc apoi matricea regresorilor notata $\mathbf{X}$:

```
X<-cbind(Sex,Mama,Tata)
```

Este o matrice de dimensiuni $N \times 3$ pentru ca am decis sa lucrez fara constanta (sau intercept). Mai jos afisez valorile pentru primii 15 indivizi ai populatiei din cei 1000.

```
> X[1:15,]
      Sex     Mama     Tata
 [1,]   0 163.0636 181.1260
 [2,]   0 163.7654 176.1571
 [3,]   0 162.9965 168.5060
 [4,]   0 163.4570 175.7353
 [5,]   0 164.1221 163.6816
 [6,]   0 162.1172 181.5270
 [7,]   0 164.2125 167.1920
 [8,]   0 165.6779 174.3481
 [9,]   1 168.2510 173.7049
[10,]   0 160.0181 167.9392
[11,]   0 165.9872 179.6918
[12,]   0 161.3570 168.9472
[13,]   0 161.2806 177.6799
[14,]   1 165.4464 179.5513
[15,]   0 164.5393 170.1735
```

Pentru a genera valorile variabilei independente va trebui sa generez mai intii vectorul erorilor modelului denumit $\boldsymbol{\varepsilon}$:

epsilon<-rnorm(N,mean=0,sd=sqrt(0.5))

Se poate observa ca am ales sa lucrez cu o valoare a variantei egala cu $\sigma^2 = 0.5$. Acum pot sa generez valorile lui $\mathbf{y}$ dupa modelul de regresie liniara scris in forma matriciala $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

y<-X%*%vecbeta+epsilon

Primele 15 valori ale lui **y**:

```
> y[1:15]
 [1] 125.62495 117.26511 105.33424 116.97726  97.16894 125.75516 102.74643
 [8] 112.56087 110.87157 107.27710 120.72043 108.19114 120.65718 120.43033
[15] 106.44341
```

Valorile minima si maxima a inaltimilor:

```
> min(y);max(y)
80.27486
129.4643
```

Asadar avind in vedere caracteristicile populatiei este mai degraba vorba de elevi si nu de studenti.

In plus, pentru a ilustra multicoliniaritatea (a se vedea mai jos) vom genera si o variabila denumita Var1 in felul urmator:

Var1<-4*Tata-3*Mama+rnorm(N,mean=0,sd=0.1)

Var1 este asadar o variabila corelata si cu inaltimea tatilor si cu inaltimea mamelor caci daca ma uit la coeficientii de corelatie volorile sint mari:

```
> cor(Tata,Var1);cor(Mama,Var1)
0.8899334
-0.4890197
```

In continuare voi estima patru modele:

- Modelul 1 contine ca regresori cele trei variabile Sex, Mama si Tata
- Modelul 2 contine regresorii Sex si Mama
- Modelul 3 contine regresorii Sex si Tata
- Modelul 4 contine regresorii Sex, Mama, Tata si Var1.

Modelul 1 este cel corect pentru ca are ca regresori variabilele folosite pentru generarea lui **y**. Modelele 2 si 3 le estimam pentru a vedea ce se intimpla daca in practica omitem un regresor important. Modelul 4 este folosit pentru a vedea ce se intimpla daca se adauga o variabila suplimentara care in plus este corelata cu regresori deja existenti in model (capcana multicoliniaritatii).

Pentru a estima modelele va trebui sa prelevam un esantion. Am ales talia esantionului egala cu n=50 si un esantion aleator simplu, adica un esantion in care totii indivizii populatiei au aceeasi probabilitate de a fi selectionati. Facem aceasta cu o functie R pe care o vom studia si folosi in anul 3:

```
srsworrs<-function(N,n){
        u<-runif(N);a<-order(u)
        return((1:N)[a][1:n])
}
```

Are doua argumente: N=talia populatiei si n=talia esantionului. Rezultatul functiei este un vector care contine indicii indivizilor selectionati. Iata un rezultat al apelarii ei:

```
> ind<-srsworrs(N=1000,n=50)
> ind
 [1] 107 287 692 172 569 448 888 765  74 941 336 137 114 253 558 816 196 422 713
[20] 927 278 847 270 339 752 574 919 176 417 514  50 932 419 305 607 100 405  88
[39] 185 612 779 431 811 584 174 922 730 814 718 587
```

Asadar s-au selectat indivizii de pe pozitiile 107, 287, 692, etc...

Vrem sa studiem empiric proprietatile estimatorilor obtinuti estimind cele patru modele de mai sus si sa comparam cu teoria de la curs. Cu alte cuvinte vrem sa vedem deplasarile si variantele lor. In acest scop vom face un studiu Monte-Carlo. Adica vom estima de un numar suficient de mare (G=10000) cele 4 modele si vom calcula valorile Monte-Carlo ale mediilor si variantelor estimatorilor. Vom face aceasta cu ajutorul unui bucle *for* ca mai jos (mai multe explicatii vom da la cursul urmator):

```
n<-50;G<-10000
estbeta1<-matrix(,G,3);estbeta2<-matrix(,G,2);
estbeta3<-matrix(,G,2);estbeta4<-matrix(,G,4)
estsigma1<-numeric(G);estsigma2<-numeric(G)
estsigma3<-numeric(G);estsigma4<-numeric(G)
for(g in 1:G){
ind<-srsworrs(N=N,n=n)
ys<-y[ind];
Xs1<-X[ind,];Xs2<-X[ind,-3];
Xs3<-X[ind,-2];Xs4<-(cbind(X,Var1))[ind,]

estbeta1[g,]<-as.numeric(solve(t(Xs1)%*%Xs1)%*%t(Xs1)%*%ys)
estbeta2[g,]<-as.numeric(solve(t(Xs2)%*%Xs2)%*%t(Xs2)%*%ys)
estbeta3[g,]<-as.numeric(solve(t(Xs3)%*%Xs3)%*%t(Xs3)%*%ys)
estbeta4[g,]<-as.numeric(solve(t(Xs4)%*%Xs4)%*%t(Xs4)%*%ys)

estys1<-Xs1%*%estbeta1[g,]
estys2<-Xs2%*%estbeta2[g,]
estys3<-Xs3%*%estbeta3[g,]
estys4<-Xs4%*%estbeta4[g,]

estrez1<-(ys-estys1)
estrez2<-(ys-estys2)
estrez3<-(ys-estys3)
estrez4<-(ys-estys4)

estsigma1[g]<-sum(estrez1^2)/n
estsigma2[g]<-sum(estrez2^2)/n
estsigma3[g]<-sum(estrez3^2)/n
estsigma4[g]<-sum(estrez4^2)/n
}
```

Acum cer mediile estimatorilor vectorilor $\boldsymbol{\beta}$ pentru fiecare model:

```
> colMeans(estbeta1);
[1]  0.4886631 -0.8977858  1.4979447
> colMeans(estbeta2);colMeans(estbeta3);
[1] 0.6949750 0.6429111
[1] 0.4220353 0.6261686
> colMeans(estbeta4)
[1] 0.49789680 0.03689298 0.25177772 0.31147578
```

Se poate observa ca in cazul Modelului 1 (cel corect) analiza Monte-Carlo confirma teoria: cind modelul este corect estimatorul $\hat{\boldsymbol{\beta}}$ este nedeplasat.

Pentru Modelul 2, in care am omis un regresor important (Inaltimea Tatalui) media coeficientului Sexului este 0.694 iar media coeficientului Inaltimii Mamei este 0.642 ceea ce arata o deplasare consistenta (mai ales la inaltimea mamei unde coeficientul considerat era negativ egal cu -0.9). O astfel de schimbare de semn poate duce la concluzii eronate. In cazul de fata am deduce ca inaltimea mamei inseamna o inaltime mare a copilului ori dupa cum am generat populatia lucrurile stau pe dos. Aceelasi observatii si pentru Modelul 3. Concluzia este ca in cazul omiterii unui regresor sau regresori importanti estimarea este deplasata iar concluziile pot fi contrare realitatii.

In cazul Modelului 4 (multicoliniaritate) consecintele sint la fel de rele. Estimatorii (cu exceptia coeficientului Sexului) sint deplasati.In cazul Inaltimii Tatalui deplasarea este foarte mare iar in cazul Inaltimii Mamei avem din nou o schimbare de semn.

Pot calcula apoi variantele estimatorilor:

```
> colMeans((estbeta1-colMeans(estbeta1))^2);
[1] 1.058921 2.554279 2.252981
>colMeans((estbeta2-colMeans(estbeta2))^2);
colMeans((estbeta3-colMeans(estbeta3))^2);
[1] 9.256664102 0.001412139
[1] 3.19179457 0.02086518
> colMeans((estbeta4-colMeans(estbeta4))^2)
[1]  0.1538694  9.6892284 17.0089031  1.1027406
```

Variantele pot fi mult mai mici insa in cazul estimatorilor deplasati cu deplasare mare varianta nu mai este potrivita pentru a masura precizia. In cazul asta se considera Eroarea Patratica Medie (*EPM* sau *MSE* in engleza de la mean squared error). $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = B(\hat{\theta})^2 + V(\hat{\theta})$.

Astea s-ar obtine in felul urmator:

```
> (colMeans(estbeta1)-vecbeta)^2+colMeans((estbeta1-colMeans(estbeta1))^2);
[1] 1.059050 2.554284 2.252985

> (colMeans(estbeta2)-vecbeta[-3])^2+colMeans((estbeta2-colMeans(estbeta2))^2)
[1] 9.294679 2.381987

> (colMeans(estbeta3)-vecbeta[-2])^2+colMeans((estbeta3-colMeans(estbeta3))^2)
[1] 3.1978731 0.7844465

> (colMeans(estbeta4)[1:3]-vecbeta)^2+colMeans((estbeta4-colMeans(estbeta4))^2)[1:3]
[1]  0.1538738 10.5669969 18.5669620
```

Ne putem uita apoi la estimatorii lui σ^2

```
> mean(estsigma1)
[1] 0.4655294
> mean(estsigma2);mean(estsigma3)
[1] 68.72568
[1] 23.62318
> mean(estsigma4);
[1] 0.4541823
```

In cazul modelelor 2 si 3 estimatorii au deplasari foarte mari. In cazul modelelor 1 si 4 sint aproape nedeplasati cum indica si teoria. Pt a obtine estimatorii exact nedeplasati se modifica putin estimatorii (n=50 nu este foarte mare):

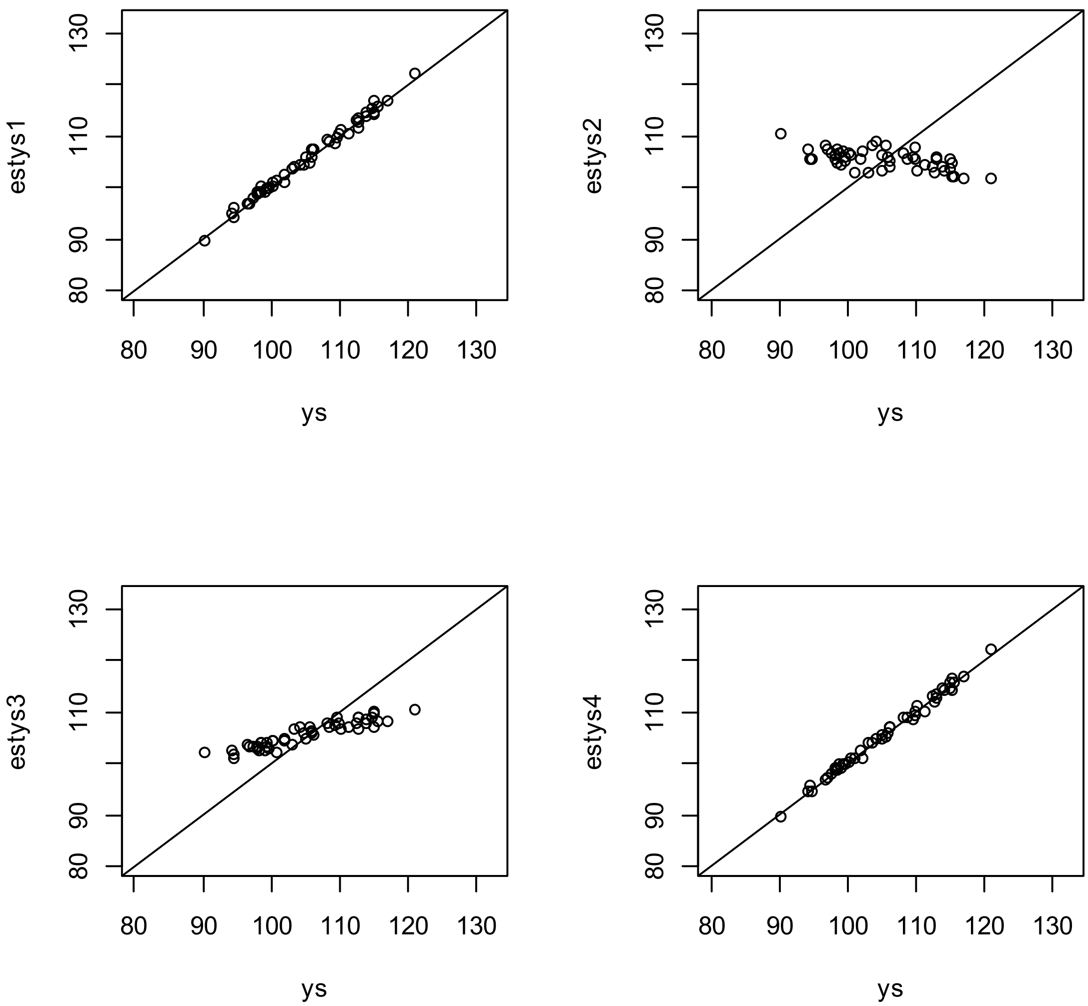
```
> mean(estsigma1*n/(n-3))
[1] 0.495244
> mean(estsigma2*n/(n-2));mean(estsigma3*n/(n-2))
[1] 71.58925
[1] 24.60748
> mean(estsigma4*n/(n-4))
[1] 0.4936764
```

Vreau acum sa vizualizez cum se pozitioneaza $\hat{y}_i$ dat de fiecare model in raport cu observatiile y_i. Fac aceasta pt un esantion din cele G=10000 selectionate, si anume pt ultimul. Pentru a facilita comparatia am trasat si prima bisectoare. Liniile R care faca asta sint urmatoarele

```
par(mfrow=c(2,2))
plot(ys,estys1,xlim=c(min(y),max(y)),ylim=c(min(y),max(y)));abline(0,1)
plot(ys,estys2,xlim=c(min(y),max(y)),ylim=c(min(y),max(y)));abline(0,1)
plot(ys,estys3,xlim=c(min(y),max(y)),ylim=c(min(y),max(y)));abline(0,1)
plot(ys,estys4,xlim=c(min(y),max(y)),ylim=c(min(y),max(y)));abline(0,1)
```

Modelul 1 (cel corect) furnizeaza estimatii foarte bune pentru y_i. In cazul modelului 4, chiar daca coeficientii modelului erau estimati deplasat, y_i este prezis la fel de bine ca in cazul modelului 1. Modelele 2 si 3 pot da estimatii deplasate pentru y_i mai ales in cazul elevilor pitici sau mari (departe de inaltimea medie).

Fig 1



Revenim acum la Modelul 1 si la un singur esantion (cu ale cuvinte iesim din logica studiului Monte-Carlo si ne plasam in situatia cu care ne confruntam in practica). Rezultatele estimarii modelului pornind de la un esantion de talie *n*=50 le-am pus in tabelul 1:

```
> ind<-srsworrs(N=N,n=n)
> ys<-y[ind];Xs1<-X[ind,]
> estbeta1<-as.numeric(solve(t(Xs1)%*%Xs1)%*%t(Xs1)%*%ys);
> estys1<-Xs1%*%estbeta1;
> estrez1<-(ys-estys1);
> estsigma1<-sum(estrez1^2)/n;
> estbeta1
[1]  0.5764441 -0.9025569  1.5040127
> estsigma1
[1] 0.531966
```

Regresor	Estimatie Coeficient	Eroare Standard	Statistica Test Normal	Statistica TestWald (p-valori)	Statistica TestRaport deVerosimilitate (p-valori)
Sex	0.576	0.274	2.096	4.39 (0.037)	2.105 (0.146)
Inaltime Mama	-0.902	0.020	-44.39	1971.32 (<0.001)	92.48 (<0.001)
Inaltime Tata	1.504	0.019	76.41	5838.83 (<0.001)	119.21 (<0.001)

Estimatiile coeficientilor au erori standard care se obtin de pe diagonala matricii de varianta-covarianta $\hat{V}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1}$. Matricea se obtine asa:

```
> estsigma1*solve(t(Xs1)%*%Xs1)

              Sex          Mama          Tata
Sex    0.075617766 -0.0011971920  0.0010819628
Mama  -0.001197192  0.0004132297 -0.0003998824
Tata   0.001081963 -0.0003998824  0.0003874156
```

Cu estimatiile coeficientilor si cu estimatiile erorilor standard putem testa pt fiecare coeficient ipoteza daca acesta este 0 (adica regresorul este nesemnificativ statistic) versus ipoteza ca este diferit de zero (adica regresorul este semnificativ statistic).

Aceasta se poate face fie cu ajutorul cuantilelor distribuiei statisticii de test sub ipoteza nula fie calculind p-valoarea testului si comparind-o cu pragul $\alpha = 0.05$.

In cazul testului normal pot sa folosesc cuantila de ordin 0.975 care este egala cu qnorm(0.975)=1.959964. In cazul testului Wald, p-valoarea este probabilititatea de a obtine o valoarea a statisticii de test mai extrem decit ce s-a obtinut. In cazul nostru, pentru coeficientul regresorului Sex (si similar si pt celelalte) este probabilitatea ca o variabila distribuita $\chi^2(1)$ sa ia o valoare mai mare decit 4.39. Cu R asta se calculeaza asa:

```
> 1-pchisq(df=1,4.34)
[1] 0.03722692
```

La fel si pentru ceilalti doi regresori:

```
> 1-pchisq(df=1,1971.32)
[1] 0
> 1-pchisq(df=1,5838.83)
[1] 0
```

In cazul ultimilor doi probabilitatile sint extrem de mici ceea ce arata semnificativitatea foarte mare a inaltimilor parintilor. Si Sexul este semnificativa deoarece p-valoarea=0.037 este mai mica decit pragul 0.05 in care caz se respinge ipoteza nula. Dar gradul de semnificativitate este mult ai mic.

Mai exista si un alt test (likelyhood ratio test) sau testul raportului de verosimilitate. Acesta foloseste estimatia logului functiei de verosimilitate. Pentru modelul 1 aceasta se estimeaza dupa formula:

$$\log(\hat{L}) = -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}\log(2\pi)$$

Cu R aceasta se calculeaza asa:

```
> logverosimilitate1<--sum(estrez1^2)/estsigma1/2-n/2*log(estsigma1)-n/2*log(2*pi)
> logverosimilitate1
[1] -55.16753
```

Pentru a calcula statistica de test necesara testarii semnificativitatii regresorului Sex va trebui reestimat modelul fara regresorul Sex si calculat logul verosimilitatii modelului fara Sex:

```
> Xs11<-X[ind,-1]
> estbeta11<-as.numeric(solve(t(Xs11)%*%Xs11)%*%t(Xs11)%*%ys);
> estys11<-Xs11%*%estbeta11;
> estrez11<-(ys-estys11);
> estsigma11<-sum(estrez11^2)/n;
logverosimilitate11<--sum(estrez11^2)/estsigma11/2-n/2*log(estsigma11)-n/2*log(2*pi)
> logverosimilitate11
[1] -57.27345
```

Se poate observa ca in cazul in care adaugam regresori la un model (in cazul asta Sexul adaugat modelului fara Sex pt a obtine modelul 1) logverosimilitatea (si deci si verosimilitatea) creste. In cazul nostru cresterea este egala cu 2.105:

```
> logverosimilitate1-logverosimilitate11
[1] 2.105913
```

Intrebarea este daca 2.105 este suficient de mare astfel incit sa merite adaugarea regresorului Sex la model. Semnificativitatea lui 2.105 se testeaza din nou cu ajutorul distributiei $\chi^2(1)$:

```
a<-logverosimilitate1-logverosimilitate11
1-pchisq(df=1,a)
[1] 0.1467308
```

Se obtine o p-valoare egala cu 0.146 care este mai mare decit pragul 0.05. In cazul Sexului testul raportului de verosimilitate arata mai degraba ca modelul cu Sex si fara Sex au cam acelasi fit.

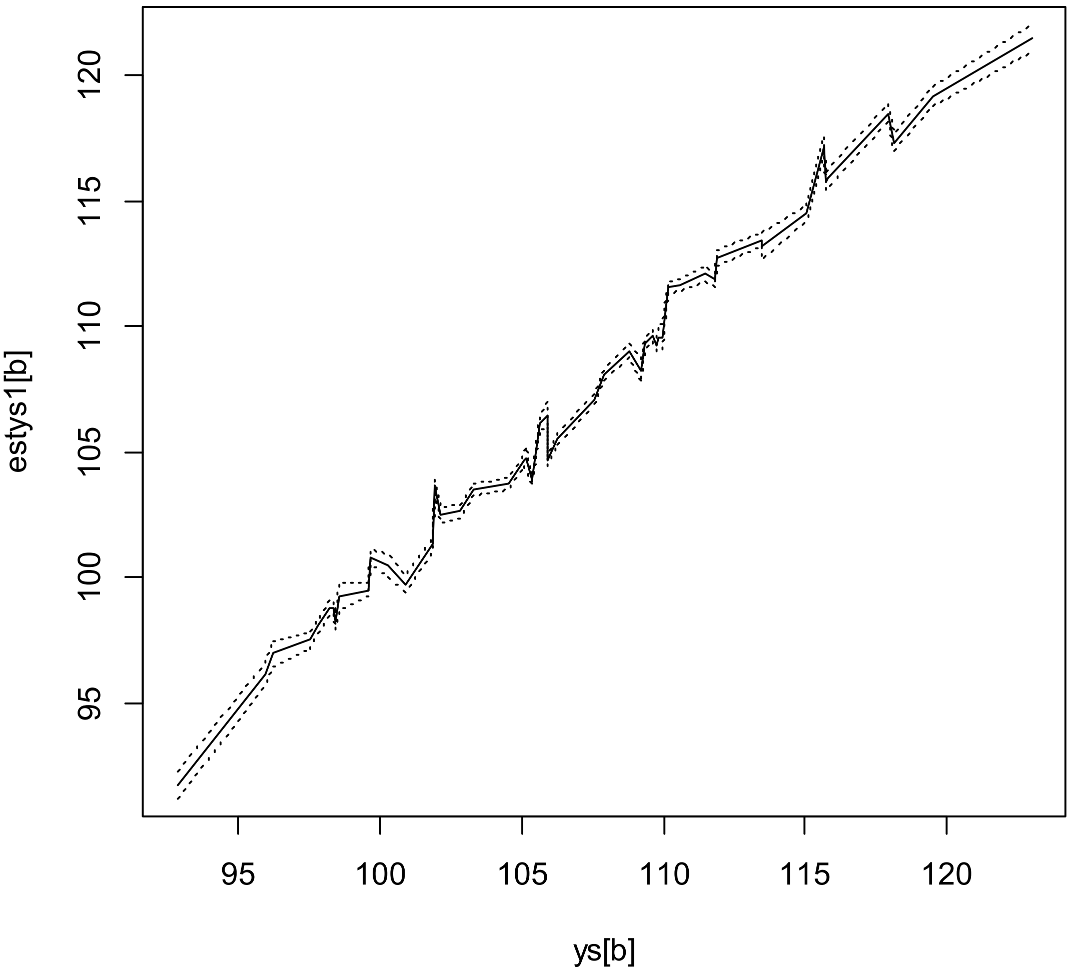
Refac calculul pentru ceilalti doi regresori si obtin valorile din tabelul 1. Ca si testul Wald, si testul raportului de verosimilitate ii indica ca fiind foarte semnficativi.

Restul componentelor matricii $\hat{V}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1}$ se folosesc pentru a calcula preciziile estimatiilor $\hat{y}_i$. Aceste estimatii le pun intr-un vector notat $\hat{\mathbf{y}}$ si care este egal cu $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ Matricea lui de varianta-covarianta va fi estimata prin $\hat{V}(\hat{\mathbf{y}}) = \mathbf{X}\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{X}^t = \hat{\sigma}^2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. Cu elementele de pe diagonala sa principala pot construi intervale de incredere de nivele de incredere 0.95 pt fiecare $\hat{y}_i$ si le pot reprezenta:

```
  estys1<-Xs1%*%estbeta1;
> a<-diag(estsigma1*Xs1%*%solve(t(Xs1)%*%Xs1)%*%t(Xs1))
> L1<-estys1-1.96*sqrt(a);L2<-estys1+1.96*sqrt(a)
> b<-order(ys)
> plot(ys[b],estys1[b],type="l")
> lines(ys[b],L1[b],lty=3)
> lines(ys[b],L2[b],lty=3)
>
```

Iar figura care rezulta este urmatoarea:

Fig 2



Estimatiile sint precise pentru ca intervalele de incredere sint inguste.