Optimització al Compilador GCC

Jordi Romero Morcillo, Pau Garcia Rodriguez i Alexis Rico Carreto

RESUM

Objectiu: Veure si el nombre de línies en llenguatge de baix nivell és el mateix si el compilador optimitzar el codi i veure si hi ha alguna relació entre el nombre de línies en C++ i el nombre de línies de la seva traducció a baix nivell.

Mètodes: Hem agafat programes d'altres assignatures i els hem compilat per obtenir les dades del nostre estudi.

Resultats: Al resoldre la nostra prova d'hipòtesi hem vist que les premisses dels nostres objectius es complien i hem obtingut els càlculs suficients per arribar a una conclusió.

Discussió: A partir dels resultats obtinguts hem vist que el nombre de línies en llenguatge de baix nivell no és igual amb o sense optimització al compilador. També hem pogut estimar l'equació lineal que relaciona el nombre de línies en C++ i el nombre de línies en baix nivell d'un programa.

INTRODUCCIÓ

En l'àmbit de la informàtica, el compilador és clau per l'execució d'un programa. El compilador determina la eficiència i rapidesa dels programes. A més, el compilador permet algunes opcions (o flags en anglès) per tal que la traducció del nostre programa sigui òptima en termes de rendiment, espai, temps, etc.

En aquest estudi hem plantejat dos **objectius** principals:

1. Veure si el nombre de línies en llenguatge de baix nivell (assemblador) dels programes de la mostra és el mateix si li demanem al compilador que optimitzi el nostre codi, o si ho deixem sense la optimització (per defecte).

2. Veure si hi ha una relació entre el nombre de línies en C++ i el nombre de línies en baix nivell (assemblador).

MATERIAL I MÈTODES

El material principal per realitzar l'estudi, ha sigut el compilador GCC Linux x86_64, dues de les opcions del compilador d'optimització i els scripts que hi ha als annexes. El script "copyScript.sh" extreu els programes candidats, i el script "script.sh" executa i retorna els resultats.

Les opcions o "flags" utilitzats han sigut -O0 i -O2 del GCC. Per a més informació sobre aquests, i altres "flags" d'optimització podeu visitar la pagina web del GCC (https://goo.gl/ZXivGV).

A tot l'estudi ens referirem indistintament a no utilitzar optimització com posar la "flag" -O0 al compilador, i a utilitzar optimització com a afegir la "flag" -O2, ja que són equivalents.

La mostra experimental es conforma per programes de diferents assignatures d'alguns estudiants de PRO1, PRO2 i EDA. Aquests programes s'han extret del lloc web www.jutge.org.

Les dades de la mostra han sigut generades per un seguit de "scripts", elaborats per nosaltres, que recollien del compilador els resultats amb les dues "flags" ja esmentades.

Per fer l'estudi de les dades hem utilitzat el software estadístic R (v.2.13.1), de la companyia *"The R Foundation for Statistical Computing"*.

A fi d'assolir el nostre primer objectiu, hem formulat una variable categòrica X amb els valors: "Sense optimització" i "Amb optimització".

A més, hem definit les següents variables resposta:
- Y1: "Nombre de línies en assemblador sense optimització".
- Y2: "Nombre de línies en assemblador amb optimització".

Segons aquestes variables, hem procedit a realitzar l'estudi amb dades aparellades (les nostres variables actuen sobre les mateixes dades) sota aquesta premissa, hem definit la variable diferencia D que representa la diferencia per cada valor de la mostra que prenen les variables Y1 i Y2.

Durant l'estudi del nostre segon objectiu, hem agafat només una de les dues variables resposta (Y1 sense "flag" d'optimització) i una nova variable Z. Aquesta variable Z és una variable predictora i significa "Nombre de línies en C++".

MÈTODES

Per començar el nostre estudi, plantegem formalment la nostra hipòtesi i formulem la prova de la mateixa que ens ajudarà a evaluar el nostre primer objectiu. Els passos a seguir són els següents:

I. Premisses convenients

1. La variable diferència D ha de seguir una distribució Normal
2. La mostra ha de ser una mostra aleatòria simple (m.a.s.) independent
3. Efecte additiu constant a la mostra aleatòria simple

Al següent apartat demostrarem si la nostra mostra i les nostres variables compleixen aquestes premisses, i si no las compleixen que hem de fer per tal que les segueixin.

II Plantejament de la hipòtesi

- H_0 : $\mu_D = 0$
- H_1 : $\mu_D \neq 0$

D'acord amb la nostra hipòtesi farem la prova de forma bilateral.

III Estadístic i distribució de l'estadístic

Determinem l'estadístic seguint la distribució d'una t-Student:

$$\hat{t} = \frac{(\bar{D}-\mu_D)}{S_D/\sqrt{n}} = \frac{\bar{D}}{S_D/\sqrt{n}}$$

On $\bar{D}$ és la mitjana de tots els valors que pot tenir la variable D, S_D és l'estimació de la desviació estàndard de la diferència i *n* és la grandària mostral.

La nostra t-Student té *n-1* graus de llibertat on *n* és la grandària de la mostra. La distribució de l'estadístic és:

$$\hat{t} \sim t_{n-1}$$

IV Estudi del p-valor ($\alpha = 0.05$)

Si el p-valor < α, llavors podem rebutjar H_0.

V Estudi del punt crític

Si el punt crític és més petit que el valor de l'estadístic llavors podem rebutjar H_0.

VI Interval de confiança per a la diferència

Per tal d'obtenir l'interval de confiança que contindrà la mitjana de la variable diferència D amb una seguretat del 95% utilitzarem

$IC\ (\mu_D, 0.95) = \bar{D} \pm z_{0.95} \cdot \frac{S}{\sqrt{n}}$

Respecte al nostre segon objectiu de veure la relació entre línies de C++ i línies en assemblador, volem estimar la nostra mostra a un model lineal. Els passos a seguir són els següents:

I Estimació dels paràmetres

Per estimar la mostra a una equació lineal utilitzarem les variables Z i Y1 i tres nous paràmetres: el pendent ($Error_1$), la constant a l'origen ($Error_0$) i la variància (σ^2).

Aquests tres nous paràmetres són valors poblacionals i desconeguts, per tant els hem d'estimar:

$$\widehat{\beta_1} = b_1 = \frac{S_{ZY_1}}{{S_Z}^2} \qquad \widehat{\beta_0} = b_1 = \overline{Y_1} - b_0\overline{Z} \qquad \widehat{\sigma^2} = S^2 = \frac{\sum {e_1}^2}{(n-2)}$$

- S_Z, S_{Y_1} són les desviacions tipus de Z i Y1 respectivament.
- S_{ZY_1} és la covariància de Z i Y1.
- $\bar{Y}1$ i $\bar{Z}$ són les mitjanes de les variables
- e_i són els residus del model lineal de les variables $\bar{Y}1$ i $\bar{Z}$.

II Validació del model lineal

Per tal de poder validar el nostre model lineal la nostra variable resposta ha de complir les següents premisses:

1. Linealitat
2. Homoscedasticitat
3. Normalitat
4. Independència

RESULTATS

DESCRIPTIVA

La taula 1.1 mostra la mitjana, la desviació tipus, el 1r quartil, la mediana i el 3r quartil de les nostres variables resposta de tots dos objectius, és a dir, de la variable D pel primer objectiu, i de les variables Y1 i Z pel segon objectiu. Els valors de la taula 1.1 de Y1 i Z són els valors després d'aplicar la transformació logarítmica a les variables pel segon objectiu.

	Mitjana	Desviació tipus	1r Quartil	Mediana	3r Quartil
Y1	5.350	0.3799371	5.063	5.298	5.603
Z	3.103	0.4865183	2.773	3.068	3.434
D	32.30	72.52481	-19.00	33.00	66.25

Taula 1.1: Descriptiva de les variables

La figura 1 mostra la distribució de les variables que intervenen en el nostre estudi.

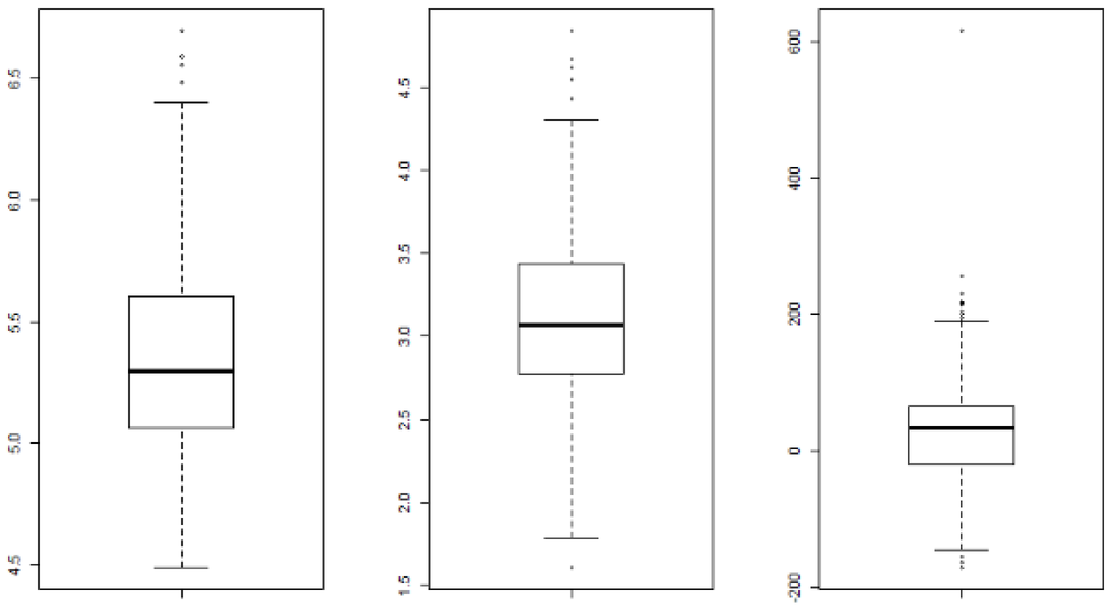

Figura 1: Boxplot de les variables Y1, Z, D

Com veiem a la figura 1 la mediana de la Y1 es troba situada més a prop del 1r quartil que del 3r quartil. Veiem que el bigoti superior és lleugerament més llarg que el bigoti inferior però suficient per indica una simetria.

Al gràfic de la variable Z, la mediana es troba situada entre el 1r i el 3r quartil. Podem veure que la longitud dels bigotis superior és gairebé la mateixa i per tant la variable Z és simètrica.

Per últim, la variable D, la mediana, es troba més a prop del 3r quartil. Els bigotis tenen la mateixa longitud i podem dir que D és simètrica.

En tots tres gràfics podem veure uns punts més enllà dels bigotis, són el que s'anomena "outliers". Aquests punts es consideren dades anòmales.

PREMISSES DEL PRIMER OBJECTIU

En aquest apartat veurem si es compleixen les premisses mencionades a l'anterior apartat.

1. **La variable diferència D ha de seguir una distribució Normal**

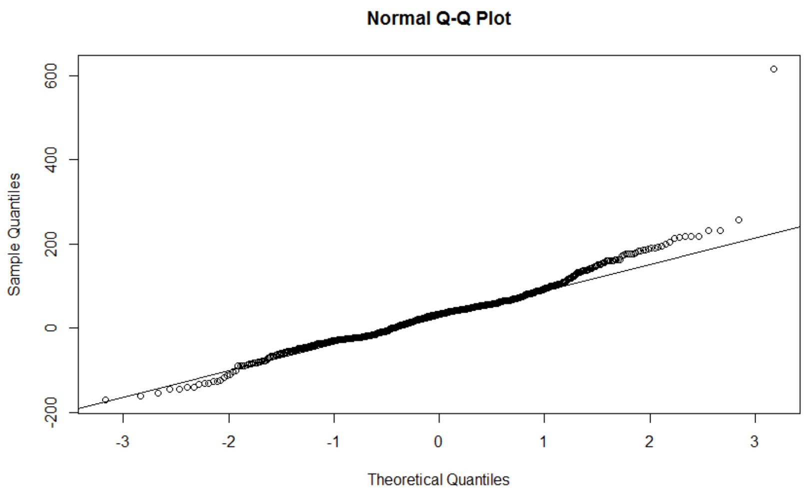

Figura 2: Qqnorm de la variable D

Com veiem a la figura 2 podem afirmar que la nostra variable D segueix una distribució Normal.

2. **La mostra ha de ser una mostra aleatòria simple (m.a.s.)**

Per tal de complir aquesta premissa la mostra ha de complir les dues condicions següents:

- Tots els elements de la població tenen la mateixa probabilitat de pertànyer a la mostra.
- Qualsevol combinació de n elements té la mateixa probabilitat de pertànyer a la mostra.

Podem afirmar que la mostra no és aleatòria simple i per tant aquesta premissa no és compleix. En el següent apartat d'aquest estudi veurem els motius pels quals no es compleix aquesta premissa.

3. Efecte additiu constant a la mostra m.a.s.

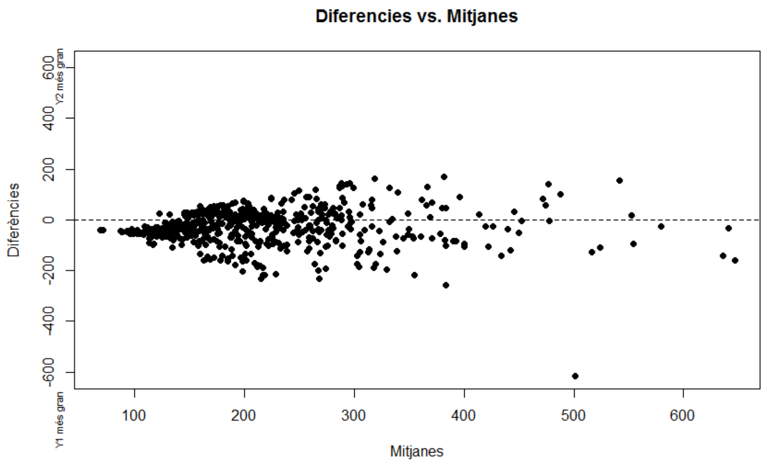

Figura 3: BlandAltman de la variable D

A la figura 3 podem veure que la mostra té un efecte additiu constant.

<u>PREMISSES DEL SEGON OBJECTIU</u>

La figura 4 mostra els 4 gràfics que demostren que les premisses de linealitat, d'homoscedasticitat, normalitat i independència és compleixen.

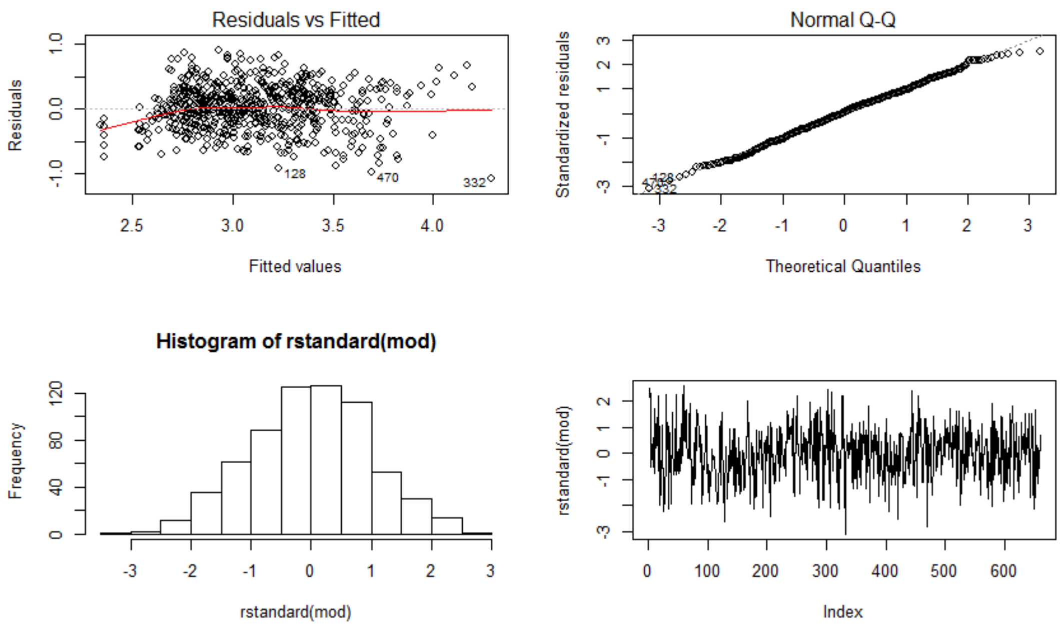

Figura 4: Gràfics e_i versus "Fitted Values", qqnorm i histograma dels residus(e_i) i ei versus ordre observacions

Amb el primer gràfic podem veure que tots els punts segueixen la recta i que la desviació estàndard és constant i per tant podem dir que la linealitat i l'homoscedasticitat és compleixen.

Al segon gràfic podem veure com totes les nostres dades segueixen la recta normal i en el tercer gràfic veiem com les nostres dades formen una campana de Gauss. Per tant, la premissa de normalitat es compleix.

A l'últim gràfic com no s'observa cap patró no tenim arguments per rebutjar la premissa d'independència.

<u>RESULTATS DEL PRIMER OBJECTIU</u>

Per provar la prova d'hipòtesi plantejada, hem de calcular el valor de l'estadístic, el p-valor i el punt crític.

Per tenir valors acurats utilitzem el RStudio:

- El valor de l'estadístic $\hat{t}$ és 11.441, amb 659 graus de llibertat.
- El p-valor és més petit que 2.2x10^(-16)
- El punt crític val 1.96357.
- L'interval de confiança del 95% és [26.75376, 37.84018]

A la discussió veurem com hem interpretat aquests resultats i a quina conclusió hem arribat respecte a la hipòtesi plantejada inicialment.

```
        Paired t-test

data:  Y1 and Y2
t = 11.441, df = 659, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 26.75376 37.84018
sample estimates:
mean of the differences
               32.29697
```

Figura 5: Resultats de la instrucció t.test.paired()

RESULTATS DEL SEGON OBJECTIU

Encara que els resultats els interpretarem a la discussió. Els valors d'estimar els paràmetres del segon objectiu $Error_1$, $Error_0$, σ^2 són:

- La variable b_1 amb valor 0.88356.
- La variable b_0 amb valor -1.62419.
- S^2 amb valor 0.3524 i S té valor 0.5936.
- R^2 amb valor 0.4761

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.07306 -0.23351  0.00792  0.24033  0.89960

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.62419    0.19381   -8.38  3.2e-16 ***
Y1           0.88356    0.03613   24.45  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3524 on 658 degrees of freedom
Multiple R-squared:  0.4761,	Adjusted R-squared:  0.4753
F-statistic:   598 on 1 and 658 DF,  p-value: < 2.2e-16
```

Figura 6: Resultats de la instrucció summary(mod)

A la figura 7 veiem la relació entre els valors de Y1 i Z

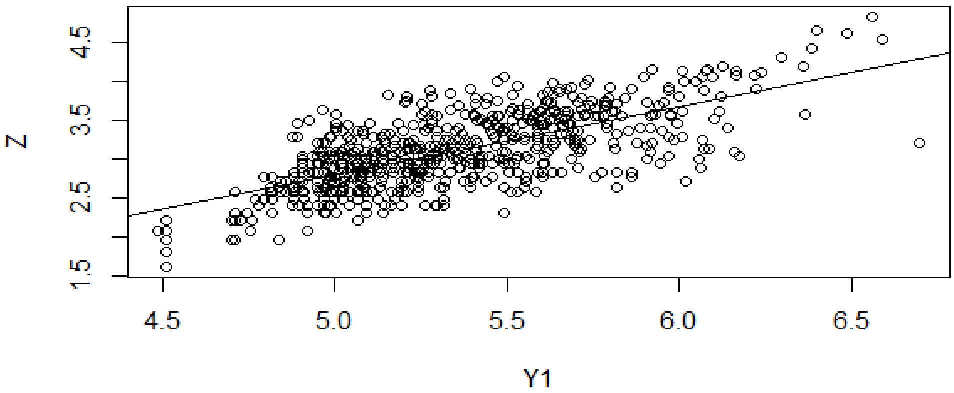

Figura 7: Y1 vs Z

DISCUSSIÓ

INTERPRETACIÓ DELS RESULTATS DEL PRIMER OBJECTIU

Com que el p-valor < Erroron α val 0,05 i el valor de l'estadístic (11.441) és més gran que el punt critic (1.96357) podem rebutjar la hipòtesi nul·la (H_0), és a dir, el nombre de línies en baix nivell sense utilitzar el "flag" O2 no és igual al nombre de línies en assemblador utilitzant el "flag" O2.

INTERPRETACIÓ DELS RESULTATS DEL SEGON OBJECTIU

La equació lineal estimada resultant és Y1 = -1.62419 + 0.88356*Z.

Interpretació de b1: Per cada línia de codi en C++, al compilar sense optimització obtenim 0.88356 línies en assemblador.

Interpretació de b0: Al compilar sense cap línia en codi C++, obtenim -1.62419 línies en assemblador.

Interpretació de S: La desviació residual val 0.5936. Podem esperar fluctuacions de 0.5936 línies d'assemblador respecte les previsions en funció de les línies de codi en c++ que ens doni el model.

Interpretació de R^2: El coeficient de determinació val 0.4761. Això implica que un 47% de les mostres es determinen per l'equació donada.

LIMITACIONS i GENERABILITAT

Limitacions. Durant el nostre estudi ens hem trobat moltes limitacions. La primera limitació va ser a l'hora de recollir les dades. La nostra mostra no compleix la premissa de ser aleatòria ja que està formada per programes d'alumnes no escollits de forma aleatòria i de assignatures arbitràries.

Una altra limitació que ens vam trobar va ser que vam haver de limitar la mostra tan sols a aquells programes que no importaven grans llibreries a fi de obtenir resultats rellevants. A l'annex hi ha els gràfics ho mostren.

L'última limitació que ens hem trobat va ser la necessitat d'aplicar logaritmes a les variables Y1 i Z per tal de complir les premisses de validació *del bloc 6*. A l'annex hi ha gràfics abans d'aplicar els logaritmes.

Generabilitat. El nostre estudi inicialment volia ser general, però degut a les limitacions ja esmentades no ho podem considerar com a genèric. A més la mostra no compleix la premissa d'aleatorietat i, per tant, els resultats obtinguts poden no ser rellevants per a altres estudis similars.

ÀNNEX

PROBLEMES AMB LES DADES ORIGINALS PEL NOSTRE PRIMER OBJECTIU

Amb les dades originals fent el gràfic BlandAltman vam observar que es produïa un efecte multiplicatiu tal i com es pot veure a la figura 1.

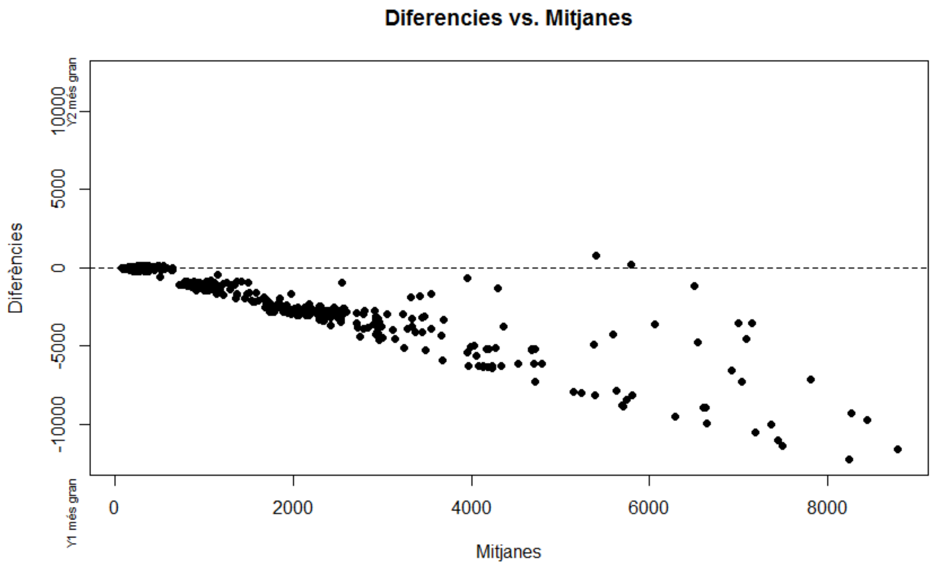

Figura 1: BlandAltman de les dades originals

Per tal de solucionar aquest problema vam aplicar logaritmes a la variable diferència. Quan vam tornar a fer el gràfic BlandAltman vam veure que teníem un efecte additiu constant però ens vam adonar que entre algunes dades hi havien un espai en blanc que separava les dades de la mostra.

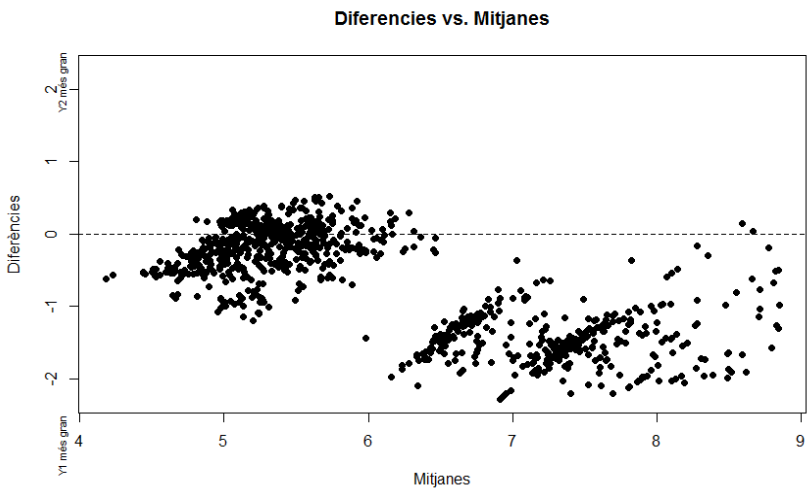

Figura 2: BlandAltman de les dades originals aplicant logaritmes

Per tal de solucionar aquest nou problema vam fer els gràfics de les variables Y1 i Y2 per separat per tal de veure on podia estar l'error. Llavors vam observar que a la Y1 es formaven dos grups i en el mig no hi havia cap dada mentre que a la Y2 no hi havia aquest problema.

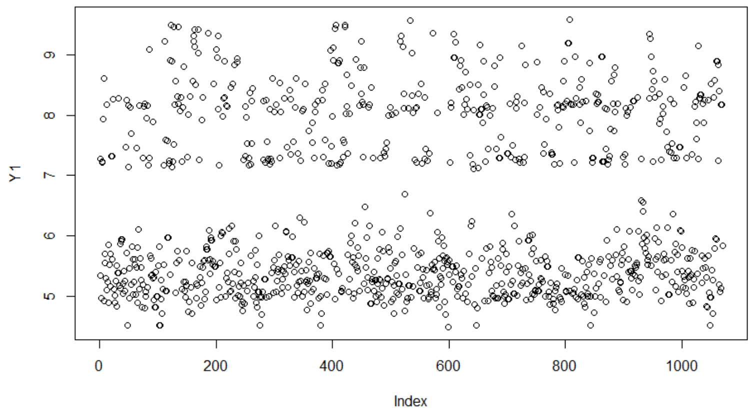

Figura 3: Gràfica dels valors de la variable Y1 aplicant logaritmes

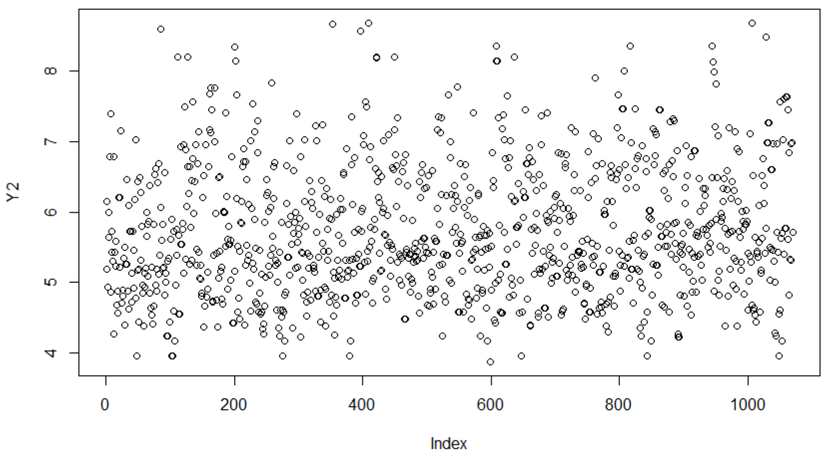

Figura 4: Gràfica dels valors de la variable Y2 aplicant logaritmes

Per tal de solucionar aquest problema vam decidir treure tots aquells programes que el nombre de línies en assemblador sense optimitzar era més gran que 1050.

PROBLEMES AMB LES DADES PEL NOSTRE SEGON OBJECTIU

Amb la variable Y1 ja modificada i la variable predictora Z vam fer l'estudi pel nostre segon objectiu. Fent les gràfiques per veure si es complien les premisses de linealitat, homoscedasticitat, normalitat i independència vam veure que la premissa d'homoscedasticitat no es complia tal i com podem veure a la figura 5 i 6.

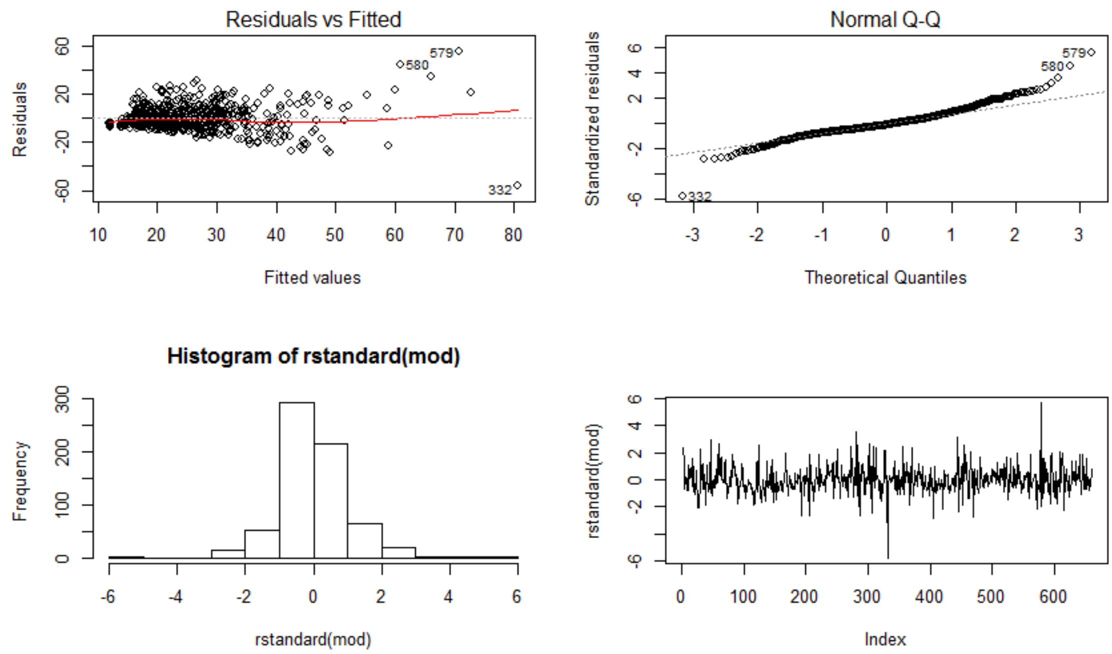

Figura 5: Gràfics amb les variables Y1 i Z

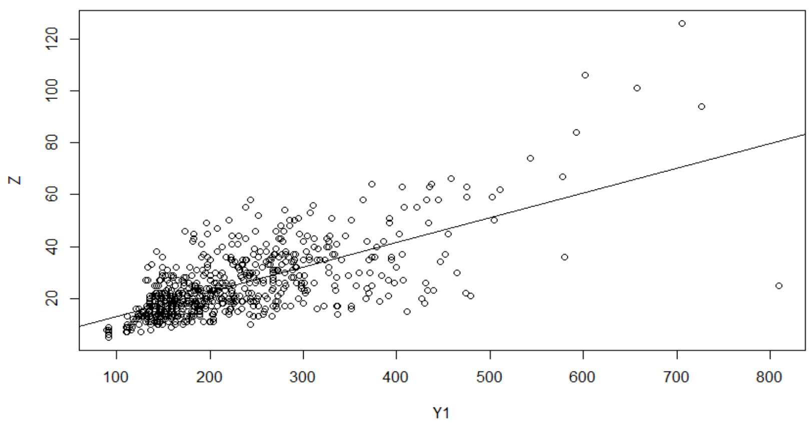

Figura 6: Gràfic Y1 vs Z

Per solucionar aquest problema vam aplicar logaritmes a totes dues variables i vam observar que ara si que es complien totes les premisses tal i com es pot veure en l'apartat de Resultats.