Project Statistics

Riwaaz Ranabhat
s0210700

Introduction

Zoals u kan zien beëindigt mijn studentnummer op 700 als laatste 3 cijfers dus de i,j,k op de R-code stemmen hiermee overeen.

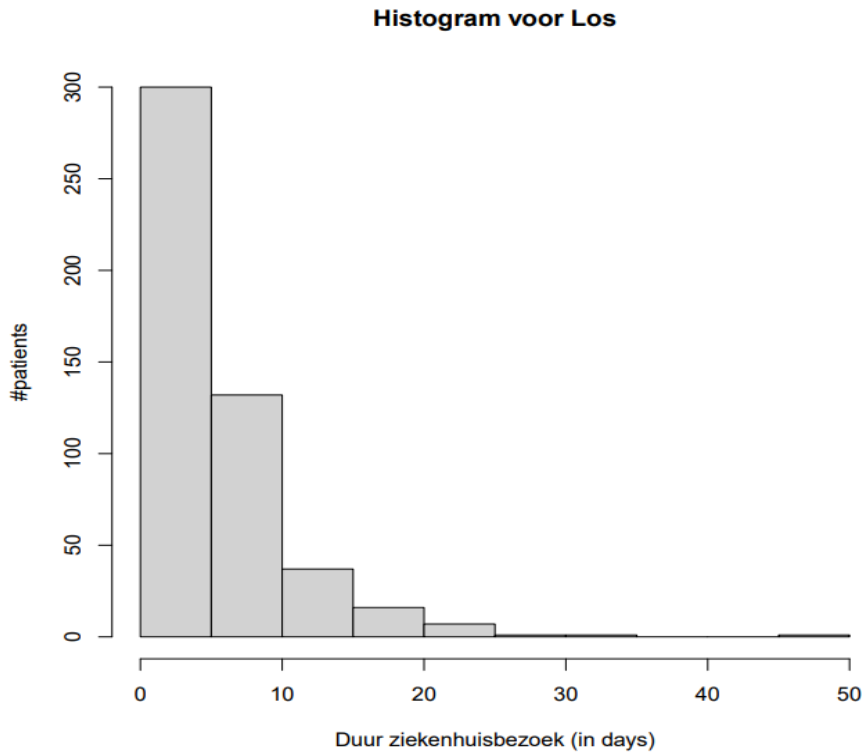
Vraag 1:

Bestudeer en bespreek de verdeling van de variabele los. Bespreek hiertoe gepaste grafische voorstellingen. Ga ook op een formele manier na of de gegevens normaal verdeeld zijn. Indien dit niet het geval is, in welke zin wijken de gegevens af van normaal verdeelde gegevens. Kan je de gegevens transformeren naar normaal verdeelde gegevens? Bespreek.

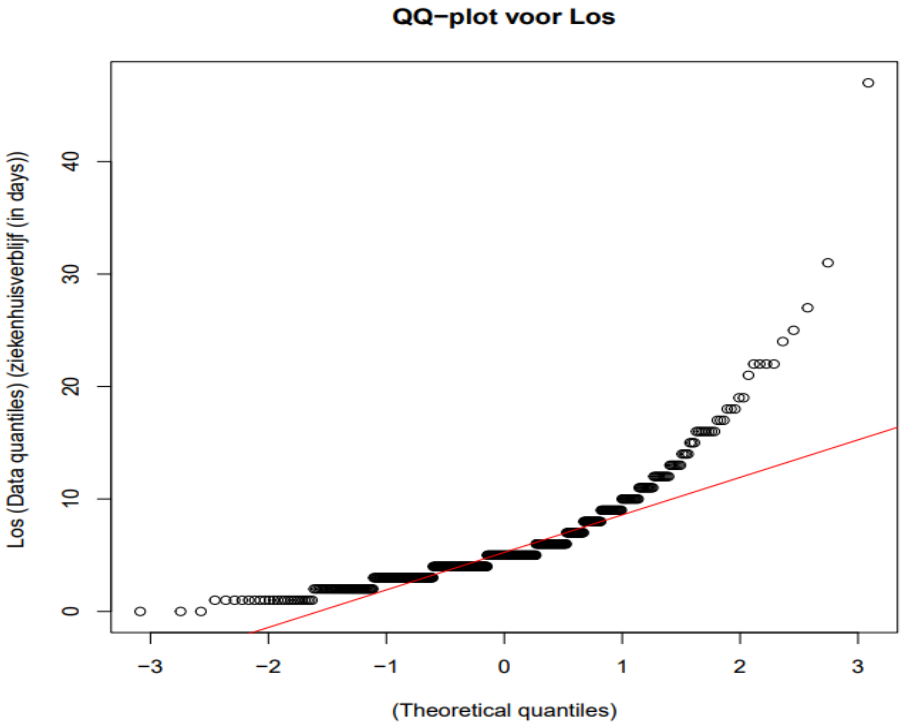
Bespreking:

- Nulhypothese: los variabele is normaal verdeeld
- Alternatieve- Hypothese: los variabele is niet normaal verdeeld

Om dit te kunnen achterhalen wilde ik graag eens starten met de spreidings kenmerken te bekijken. We merken op dat de mediaan kleinere waarde heeft dan het gemiddelde. Dit gaf me al een indicatie dat dit geen normale verdeling zou zijn. Om het verder eens te bezien heb ik dan een histogram en een qq-plot opgesteld.



Het histogram geeft ook een verdeling die rechtsscheef is. Als het een normale verdeling zou zijn dan zou dit histogram een bell shaped curve moeten hebben. Laten we eens de QQ-plot bekijken.



De QQ-plot toont ook significante afwijkingen van de rechte lijn, vooral aan de rechterkant van de plot. Dit wijst op positieve scheefheid en zwaardere uiteinden dan verwacht bij een normale verdeling. In een normale verdeling zouden de punten langs de rechte lijn vallen, wat hier niet het geval is.

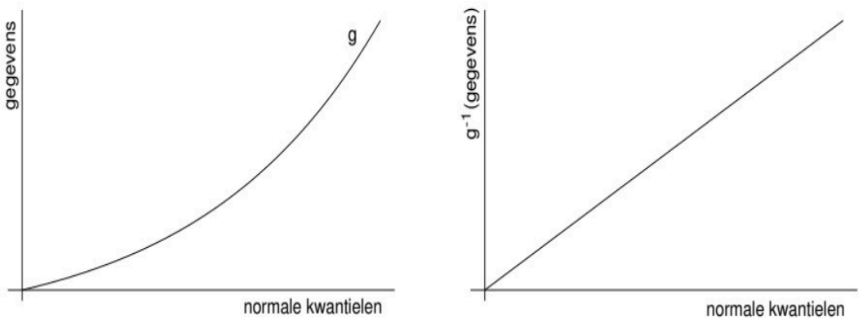
Voor de verificatie heb ik beslist om een Shapiro-Wilk test uit te voeren sinds we dit moesten doen bij onze computer zettingen. Op Zich kon ik waarschijnlijk ook een Kolmogorov-Smirnov test doen maar die test vergelijkt 2 verdelingen met elkaar terwijl Shapiro-Wilk straight ziet of je data normaal verdeeld is of niet. Dus mijn keuze was pretty straight forward.
Als resultaat kreeg ik de volgende:

OUTPUT VALUE BEFORE CHANGING:

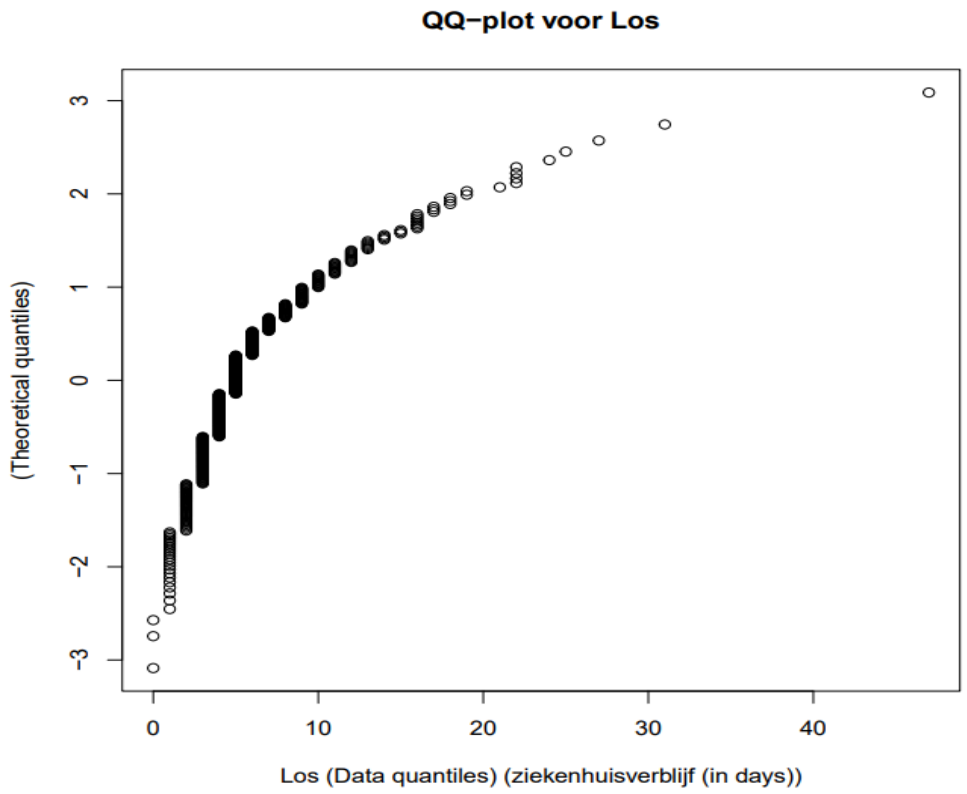
```
W = 0.76733
p-value < 2.2e-16
```

We zien dat de p-waarde veel kleiner is dan 0.05 waardoor we dan onze NulHypothese zullen afwijzen en gaan met de Alternatieve Hypothese dat de verdeling van de variabele los niet normaal verdeeld is.

De vraag is nu of we onze gegevens kunnen transformeren naar een normale verdeling. Om dit te kunnen doen heb ik eens moeten zien naar onze cursus op pagina 39 en 40 want mijn qq-plot stemde hier heel hard mee overeen.



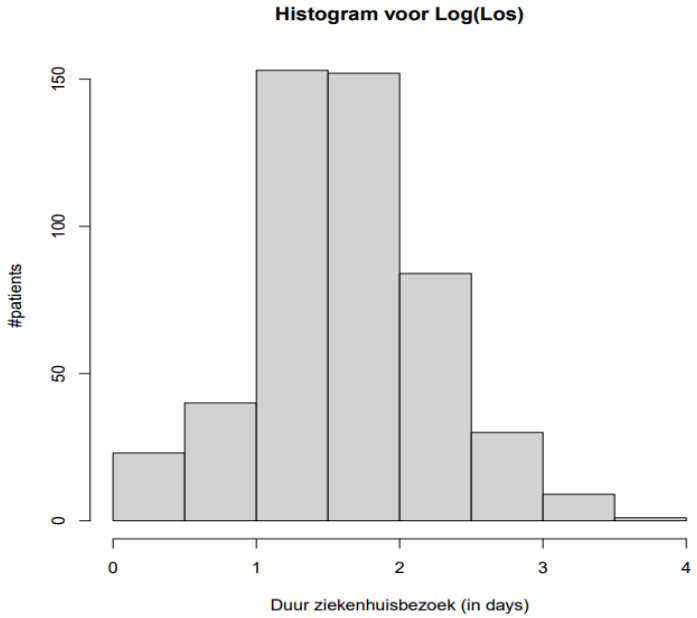
Als we hier zien dan kunnen we opmerken dat onze qq-plot overeenstemt met dit linker figuur. Om een verband te zien heb ik dan de inverse van deze functie genomen. Ik heb dan hetzelfde gedaan zoals op het cursus stond gedaan bij mijn qq-plot en ik kreeg dit:



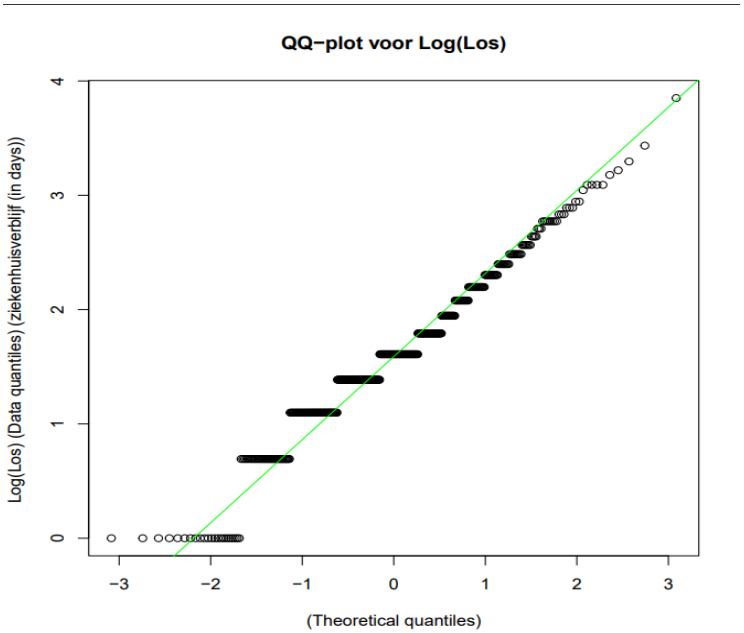
Als we naar dit zien , stemt dit grafiek heel zwaar overeen met logaritmische functie. Dus om mijn gegevens eens normaal te laten verdelen heb ik dan besloten om logaritme te nemen voor onze gegevens van los variabele.

Als we nu weer eens summary bekijken van onze los_after variable dan zien we dat de mediaan en gemiddelde min of meer same afstand van mekaar zijn dus dit wijst op meer symmetrische verdeling.

We kunne dan zelfde dingen weer terug bespreken met onze nieuwe histogram en nieuwe QQ-plot:



We zien dat onze histogram meer een bell shaped curve. Dit komt dus min of meer overeen met de normale verdeling. Laten we eens kijken naar onze QQ-plot:



We zien dat de observaties meer overeenstemmen met de groene lijn. Dit geeft ook een indicatie dat onze bewerkte gegevens door logaritme te nemen meer de normale verdeling benadert.
Als we dan hier de Shapiro test uitvoeren dan krijgen we de volgende als resultaat:

OUTPUT VALUE AFTER CHANGING:

```
W = 0.97364 met p-value = 8.87e-08
```

CONCLUSION:

We zien dat de p-value kleiner is dan 0.05 . Daarom verwerpen we de nulhypothese dat de data normaal verdeeld is en accepteren we de alternatieve hypothese dat de data niet normaal verdeeld is.

Dit resultaat is opmerkelijk omdat zowel het histogram als de QQ-plot geven aan dat de verdeling de normale verdeling min of meer benadert. Dit kan worden verklaard door het feit dat de Shapiro-Wilk test zeer gevoelig is voor kleine afwijkingen van normaliteit bij grote steekproef. Het zou dan kunnen dat we dan te maken hebben met een Type I-fout, waarbij we onterecht concluderen dat de data niet normaal verdeeld is, ondanks dat de afwijkingen klein en mogelijk onbelangrijk waren.

Vraag 2:

Ga na of er een verband is tussen het type hartinfarct en de ontslag status uit het ziekenhuis na opname. Voer een gepaste test uit.

Bespreking:

Basically, we willen graag weten of dit 2 variabelen afhankelijk zijn of onafhankelijk. We zullen hiervoor een Chi-Squared test of meer bepaald test of independence uitvoeren. As always stel ik mijn 2 hypotheses op:

- Nulhypothese: hartinfarct type en de ontslag status zijn onafhankelijk
- Alternatieve-hypothese: hartinfarct type en de ontslag status zijn niet onafhankelijk.

Vooraleer we met dit beginnen heb ik eerst een contingentietabel opgesteld waarbij we bij onze kolommen gaan bijhouden voor #levend en #dood. De rijen zullen dan de onderscheid maken voor golflengtes:

	Alive	Dead

Geen aanwezigheid van golven	318	28
Wel aanwezigheid van golven	141	11

Nu kunnen we de test hierop uitvoeren en zien wat we als output values krijgen met observed and expected table:

CHI-SQUARE OBSERVED:

	Alive	Dead
Geen aanwezigheid van golven.	318	28
Wel aanwezigheid van golven.	141	11

CHI-SQUARE EXPECTED:

	Alive	Dead
Geen aanwezigheid van golven	318.9036	27.09639
Wel aanwezigheid van golven	140.0964	11.90361

OUTPUT VALUES:

```
X-squared = 0.021371, df = 1, p-value = 0.8838
```

CONCLUSION:

We zien dat onze p-value groter is dan de significantie niveau van 0.05 dus zullen we onze Nulhypothese niet verwerpen. De variabelen dstat en mitype zijn dus onafhankelijk van mekaar.

Vraag 3:

Kan je uit de leeftijd van de patiënt het BMI voorspellen? Beantwoord deze vraag grondig en zo volledig mogelijk.

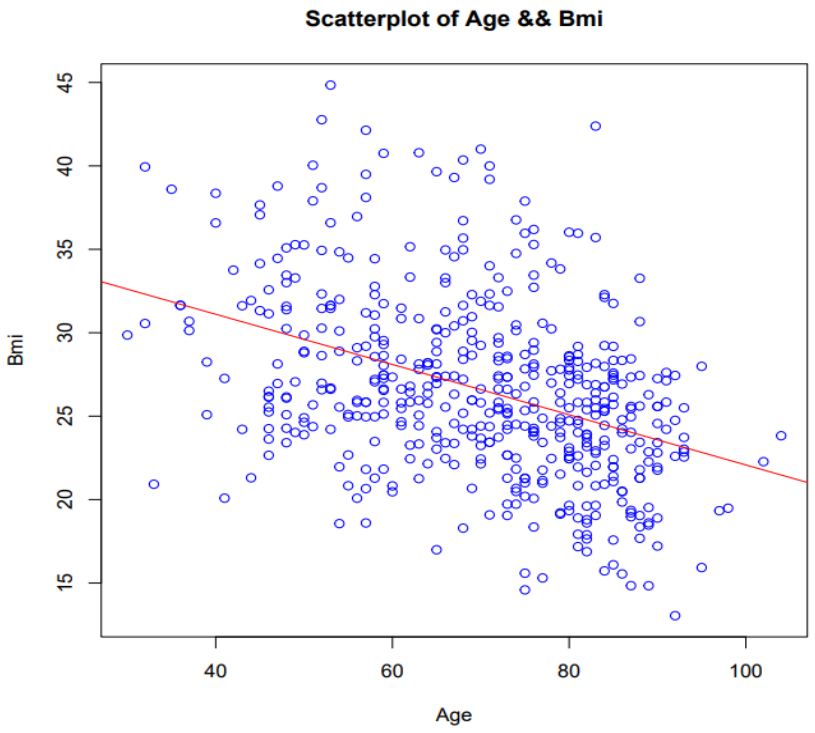
Bespreking:

Om dit te kunnen doen moeten we gebruik maken van enkelvoudige lineair regressie. Hiermee onderzoeken we op welk manier de waarde van bmi op een lineair manier voorspeld kan worden op basis van de waarde van leeftijd van de patiënt.

Vooraleer we onze regressie analyse uitvoeren moeten we eens testen of we een lineair verband hebben tussen dit 2 variables:

- Nulhypothese: Er is geen lineair verband tussen age en bmi
- Alternatieve-hypothese: Er is wel lineair verband tussen age en bmi

We beginne eerst met de scatterplot waar te nemen:



Vanuit dit waarneming kan ik al eigenlijk zeggen dat ik een negatieve lineaire
verband heb voor deze variabelen wegens de dalende trendlijn. Het lijkt wel erop
dat we een testanalyse kunnen uitvoeren maar om zeker te zijn heb ik tog eens
pearson's correlatiecoëfficiënt:

```
Pearson's product-moment correlation

data:  gegevens$age and gegevens$bmi
t = -9.7984, df = 496, p-value < 2.2e-16
alternative hypothesis: true correlation is
not equal to 0
95 percent confidence interval:
 -0.4738094 -0.3263909
sample estimates:
       cor
-0.4027084
```

we zien hierbij dat onze p-waarde weer veel kleiner is dan onze significance level
dus we mogen dan onze Nulhypothese verwerpen waar we hadden verondersteld
dat er geen lineair verband was tussen deze 2 variabelen.

REGRESSION ANALYSIS:

Alright, De bedoeling is nu dat we een vergelijking vinden waar:

$$y = \alpha + \beta x$$

waar y de waardes van bmi zijn die bepaald kunnen worden door x in te vullen die waardes voorstellen van age. We gaan eerst ofcourse onze α en β moeten
schatten.
Onze doel is nu om te zien of dat leeftijd een significant effect heeft op BMI.
Hiervoor moeten we natuurlijk ook onze 2 hypotheses opstellen:

- Nulhypothese: Leeftijd heeft geen effect op BMI. (aka. $\beta = 0$)

- Alternatieve-hypothese: Leeftijd heeft een effect op BMI.(aka. $\beta != 0$)

Om onze 2 onbekende waardes te vinden heb ik gebruik gemaakt van fitting lineair models functie. Als output kreeg ik dit :

```
Call:
lm(formula = bmi ~ age, data = gegevens)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2495  -3.4497  -0.3855   2.8184  17.7469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.12693    1.09564  33.886   <2e-16
***
age         -0.15050    0.01536  -9.798   <2e-16
***
—
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1

Residual standard error: 4.963 on 496 degrees of
freedom
Multiple R-squared:  0.1622,Adjusted R-squared:
0.1605
F-statistic: 96.01 on 1 and 496 DF,  p-value: <
2.2e-16
```

Dit betekent dat we dan de volgende equation hebben:

$$y = 37.12693 - 0.15050x$$

Deze vergelijking voorstelt dus de best fitting line. Sinds dat onze p-value weer veel

kleiner is dan onze significance level kunnen we onze nulhypothese dus verwerpen

en gaan met de alternatieve hypothese.