Laboration 2

Kungliga Tekniska Högskolan
Skolan för Elektroteknik och Datavetenskap

SF1912 - Sannolikhetsteori & Statistik

30 September 2023

Kenan Dizdarevic
kenandi@kth.se

Innehåll

1 Förberedelseuppgifter

1.1 Uppgift 1

Den stokastiska variabeln X är Rayleighfördelad och har täthetsfunktionen

$$f_X(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}.$$

$X \in Rayleigh(\sigma)$, där $\sigma = b$. Ett krav för Rayleighfördelning är oberoende, således är beräkningarna genomförda med hänsyn till att de stokastiska variablerna är oberoende.

a) ML-skattningen av b:
Sannolikhetsfunktionen för n stycken Rayleighfördelade variabler är:

$$L(b) = f_{X_1,...,X_n}(x_1, ..., x_n; b) = \{oberoende\} = f_{X_1}(x_1; b)...f_{X_n}(x_n; b)$$

$$L(b) = \prod_{i=1}^{n} \frac{X_i}{b^2} e^{-\frac{X_i^2}{2b^2}}$$

Vi skapar log-likelihood-funktionen och maximerar den med avseende på b:

$$\ell(b) = \log L(B)$$

$$\ell(b) = \log \left(\prod_{i=1}^{n} \frac{X_i}{b^2} e^{-\frac{X_i^2}{2b^2}} \right)$$

$$\ell(b) = \sum_{i=1}^{n} \log \left(\frac{X_i}{b^2} e^{-\frac{X_i^2}{2b^2}} \right)$$

$$\ell(b) = \sum_{i=1}^{n} X_i - 2n \log(b) - \frac{1}{b^2} \sum_{i=1}^{n} \frac{X_i^2}{2}$$

Maximera b genom att hitta lösningarna till $\frac{d}{db}\ell(b) = 0$

$$\frac{d}{db}\ell(b) = -2n \left(\frac{1}{b} \right) + 2 \left(\frac{1}{b^2} \right) \sum_{i=1}^{n} \frac{X_i^2}{2}$$

$$-2n \left(\frac{1}{b} \right) + 2 \left(\frac{1}{b^2} \right) \sum_{i=1}^{n} \frac{X_i^2}{2} = 0$$

$$b^*_{ML} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{X_i^2}{2}}$$

b) MK-skattningen av b:
Låt $x_1, x_2, ..., x_n$ vara observerade värden från Rayleighfördelningen med parameter b.

$$Q(b) = \sum_{i=1}^{n} (X_i - \mu(b))^2$$

Vi söker efter $E(X_i) = \mu$

$$E(X_i) = \int_0^\infty X_i \frac{X_i}{b^2} e^{-\frac{X_i^2}{2b^2}} dX = b\sqrt{\frac{\pi}{2}}$$

Detta ger oss:

$$Q(b) = \sum_{i=1}^{n} \left(X_i - b\sqrt{\frac{\pi}{2}} \right)^2$$

Vi minimerar $Q(b)$ genom att hitta lösningarna till $\frac{d}{db}Q(b) = 0$

$$\frac{d}{db}Q(b) = \frac{d}{db}\left[\sum_{i=1}^{n} \left(X_i - b\sqrt{\frac{\pi}{2}} \right)^2 \right] = 2\sum_{i=1}^{n} \left(X_i - b\sqrt{\frac{\pi}{2}} \right)\left(-\sqrt{\frac{\pi}{2}} \right)$$

$$\sum_{i=1}^{n} \left(X_i - b\sqrt{\frac{\pi}{2}} \right)\left(-\sqrt{\frac{\pi}{2}} \right) = 0$$

$$b^*_{MK} = \frac{1}{n\sqrt{\frac{\pi}{2}}} \sum_{i=1}^{n} X_i = \frac{\overline{X}}{\sqrt{\frac{\pi}{2}}}$$

$$b^*_{MK} = \overline{X}\sqrt{\frac{2}{\pi}}$$

1.2 Uppgift 2

Approximativt konfidensintervall för parametern b:
Ett approximativt konfidensintervall tar hänsyn till medelfelet av b^*_{MK}, enligt formeln $I_\theta = \theta^*_{obs} \pm D^*_{obs} t_{\alpha/2}(n-1)$. Detta kommer att tillföra t-fördelningen, ty σ är okänt. D^*_{obs} är en lämplig skattning av σ, eftersom vi har n stycken stickprov där $n \to \infty$, enligt centrala-gränsvärdessatsen.

$$D^*_{obs} = D\left[b^*\right] = \sqrt{V\left[b^*_{MK}\right]} = \sqrt{V\left[\overline{X}\sqrt{\frac{2}{\pi}}\right]}$$

$$\sqrt{\frac{2}{\pi}V\left[\overline{X}\right]} = \sqrt{\frac{2}{\pi n}V\left[X\right]} = \sqrt{\frac{2}{\pi n}\frac{4-\pi}{2}b^{*2}}$$

$$\Rightarrow D^*_{obs} = \sqrt{\frac{2}{\pi n}\frac{4-\pi}{2}b^{*2}_{MK_{obs}}}$$

$$D^*_{obs} = \sqrt{\frac{2}{\pi n}\frac{4-\pi}{2}\overline{x}^2\frac{4}{\pi^2}} = \sqrt{\frac{2\overline{x}^2(4-\pi)}{\pi^2 n}}$$

$\theta^*_{obs} = b^*_{MK}$ och $t_{\frac{\alpha}{2}}(n-1)$ går att hitta i tabell 1. Således blir vårt konfidensintervall

$$\left(b^*_{MK} - \sqrt{\frac{2\overline{x}^2(4-\pi)}{\pi^2 n}} t_{\frac{\alpha}{2}}(n-1) < b < b^*_{MK} + \sqrt{\frac{2\overline{x}^2(4-\pi)}{\pi^2 n}} t_{\frac{\alpha}{2}}(n-1) \right)$$

1.3 Uppgift 3

Linjär regression är en statistisk metod för att modellera sambandet mellan en beroende variabel och en eller flera oberoende variabler till en funktion. Målet är att hitta de bäst anpassade koefficienterna till en linjär ekvation givet den observerade datan.

Vi skall med kommandot `regress` skatta parametrarna i modellen $w = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k$. I detta fall är w den log-transformerade beroende variabeln, x_k är oberoende, β_0 är skärningen med Y-axeln, β_1 är lutningen. I denna modell antas ε_k vara en stokastisk variabel som är normalfördelad, den beskriver avvikelsen mellan det praktiska och teoretiska värdet, residualen. Detta är i sin tur en enkel linjär regression.

Koden som användes i matlab för linjär regression är:

```
1    [beta_hat, stdRes, ~, ~] = regress(w, X);
```

X är en matris där den första kolumnen innehåller ettor och den andra kolumnen innehåller datapunkternas X-värden. Detta returnerar i sin tur det uppskattade β och resiudalen ε_k (standardiserade resiudalerna). Värdena för β_i returneras i form av en vektor.

2 Laborationsuppgifter

2.1 Problem 1 - Simulering av konfidensintervall

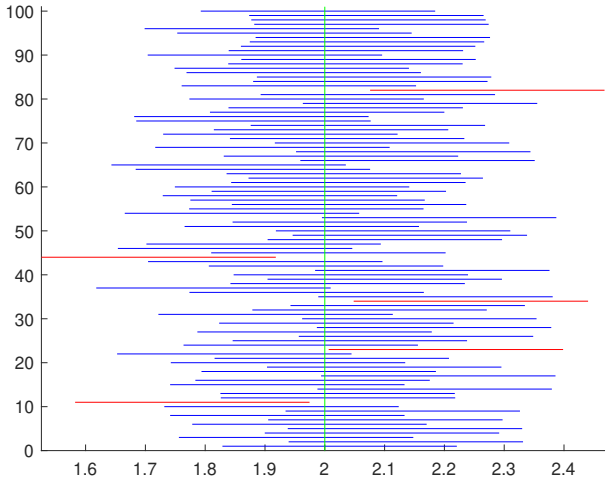
2.1.1 Problem 1.1

Vi erhåller 100 konfidensintervall. Antal konfidensintervall som kan förväntas innehålla det sanna värdet på μ är 95 intervall, eftersom $1 - \alpha = 0.95$. Resultatet kan dock variera på grund av slumpmässiga variationer i mätningarna. En konfidensgrad på 95% garanterar inte att exakt 95 av 100 intervall kommer innehålla det sanna värdet av μ. Medelvärdet borde ligga relativt nära 95.

2.1.2 Problem 1.2

De horisontella strecken visar konfidensintervallen. De blåa vertikala strecken visar konfidensintervallen som innehåller μ. De röda vertikala strecken visar konfidensintervallen som inte innehåller μ.

Det grön vertikala strecket är väntevärdet μ. Figuren nedan presenterar simulering av konfidensintervall, i detta fall har vi 5 konfidensintervall som ligger utanför väntevärdet μ. Således innehåller 95 av intervallen det sanna värdet på μ.

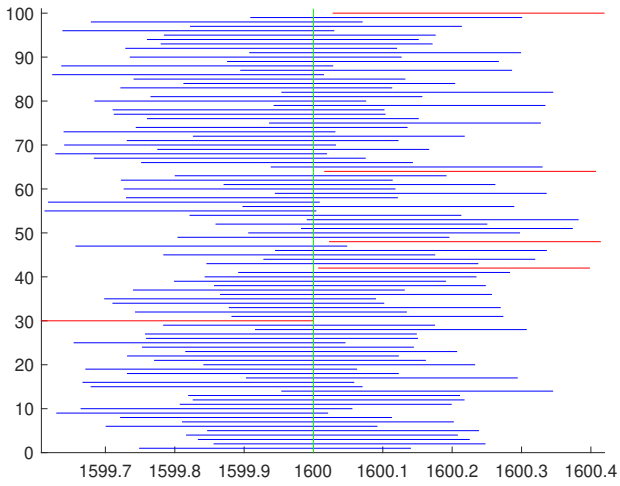


Figur 1: Simulering av konfidensintervall

2.1.3 Problem 1.3

Vi skall nu variera parametrarna μ, σ, n och α.

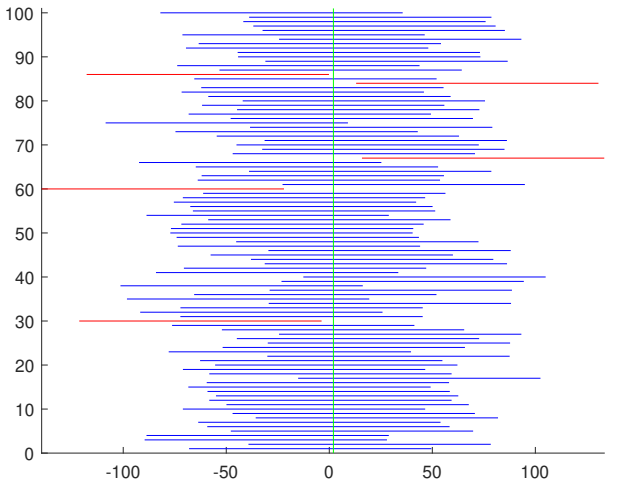
Variationen av μ presenteras i grafen nedan. Enda skillnaden från den första simuleringen är att



Figur 2: Simulering av konfidensintervall med $\mu = 1600$

grafen är förskjuten till höger. Detta är en direkt konsekvens av det nya värdet på μ. Väntevärdet förflyttar hela grafen.

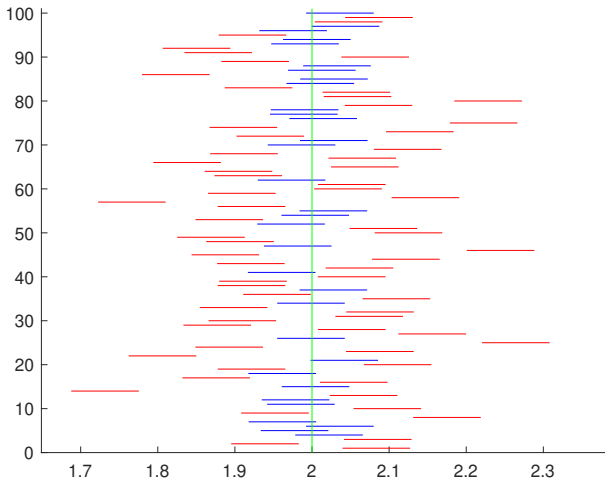
När vi ändrar på σ ser vi att vår konfidensgrad förblir den samma, men att standardavvikelsen blir större. Ett mindre σ ger en mindre standardavvikelse. Detta i sin tur innebär att konfidensintervallen antingen blir smalare eller bredare. Stora σ gör intervallen bredare och små σ gör intervallen smalare.



Figur 3: Simulering av konfidensintervall med $\sigma = 300$

Vid ökning av n skall konfidensintervallen bli smalare. Detta beror på att konfidensintervallets

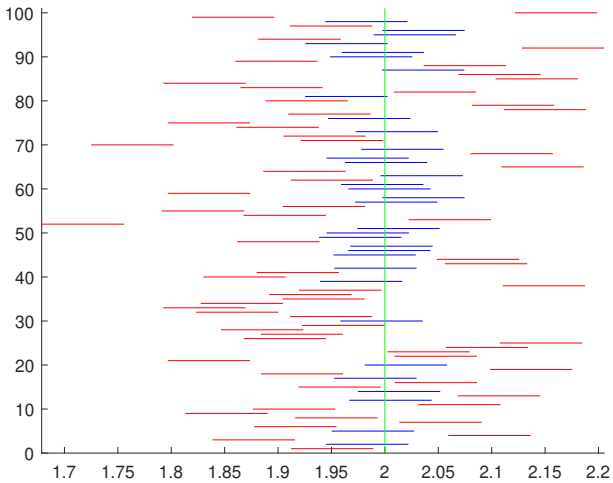
bredd är proportionellt omvänt relaterad till $\sqrt{n}$ enligt formeln $\frac{\sigma}{\sqrt{n}}$. Således skall vi få en mer exakt skattning av μ.



Figur 4: Simulering av konfidensintervall med $n = 2000$

Vi ser tydligt att konfidensintervallen blir smalare.

Vid variation av α kommer konfidensintervallet att förändras på så vis att det blir bredare om α är högre och smalare om α är mindre. Högre α ger oss större osäkerhet att μ ligger i intervallet, vi har ett brett intervall.

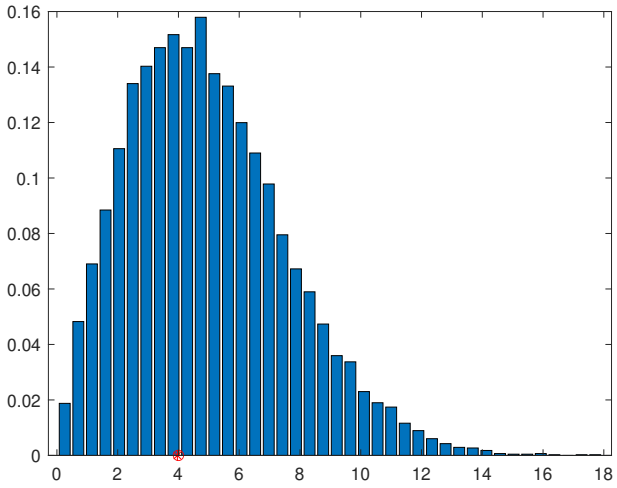


Figur 5: Simulering av konfidensintervall med $\alpha = 0.7$

2.2 Problem 2 - Maximum likelihoodskattning och minsta kvadratskattning

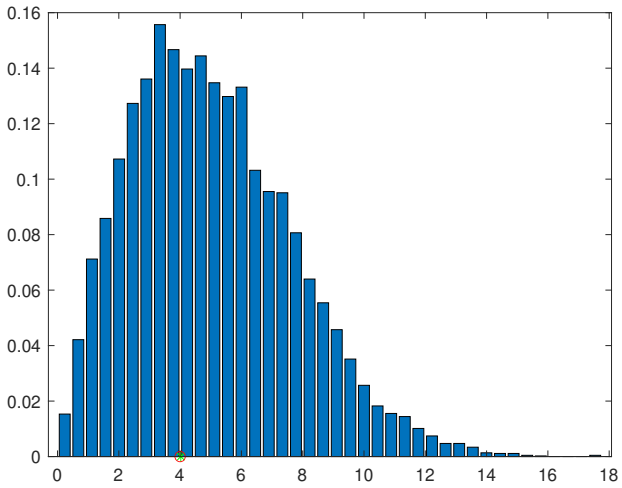
Vi skall i denna del jämföra våra skattningar för Rayleighfördelningen med dess täthetsfunktion.

2.2.1 Problem 2.1 - Simulering av ML-skattning



Figur 6: Simulering av b^*_{ML}.

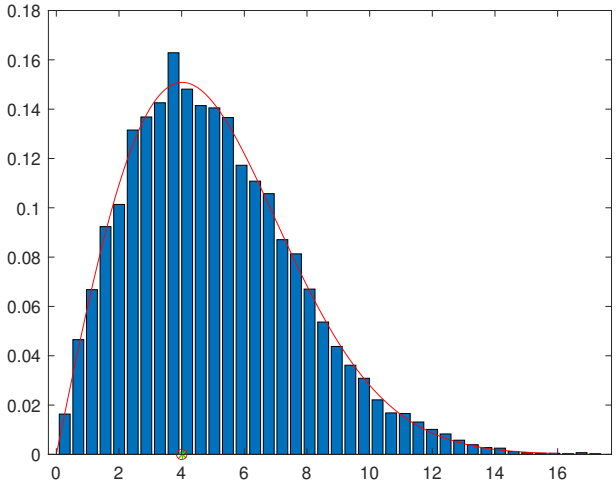
2.2.2 Problem 2.2 - Simulering av MK-skattning



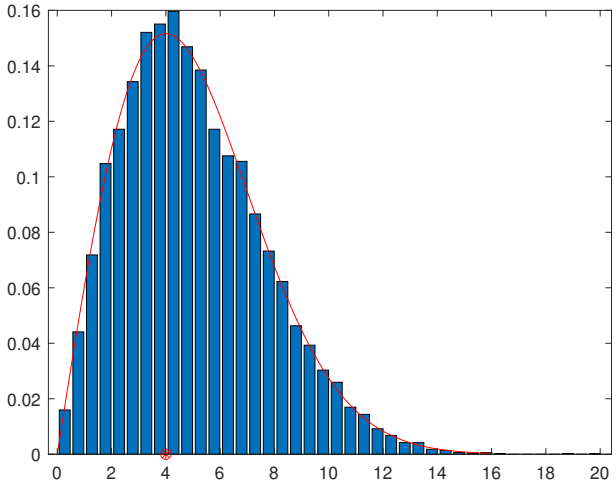
Figur 7: Simulering av b^*_{MK}.

2.2.3 Problem 2.3 - Jämförelse av skattningar

I grafen nedan jämför vi våra skattningar av b med täthetsfunktionen för en Rayleighfördelning. Våra skattningar ser bra ut, eftersom de följer täthetsfunktionen för Rayleighfördelningen.



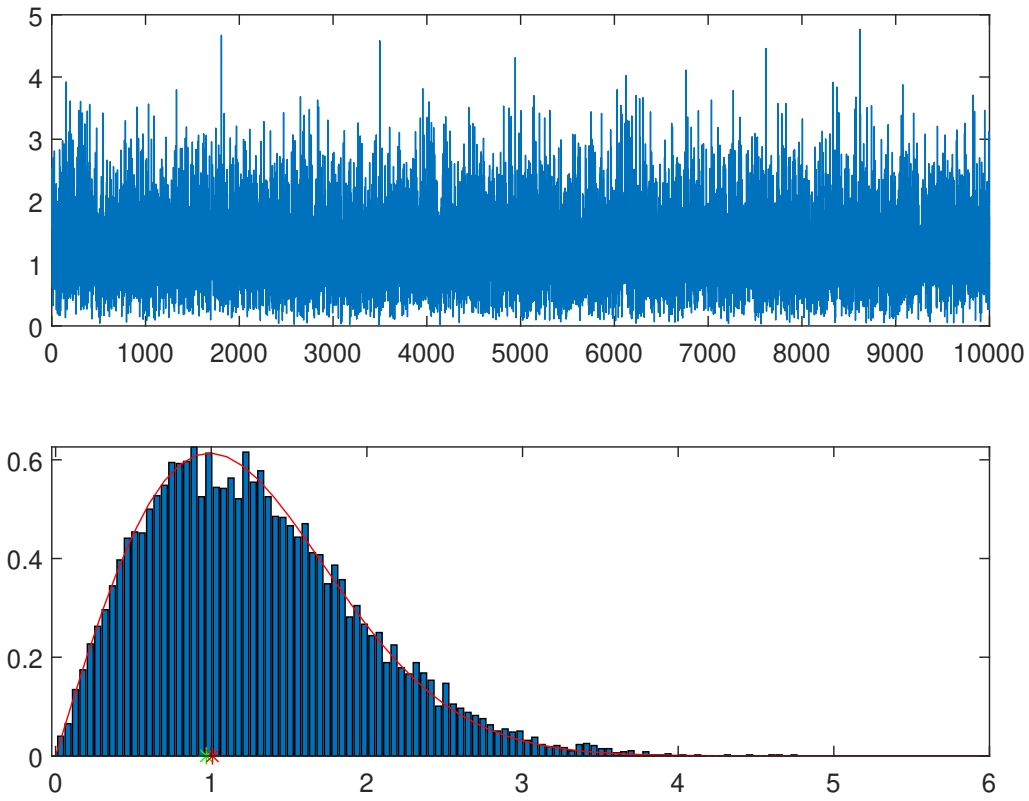
Figur 8: Simulering av täthetsfunktionen för Rayleighfördelning med b^*_{MK}.



Figur 9: Simulering av täthetsfunktionen för Rayleighfördelning med b^*_{ML}.

Båda skattningarna ser väldigt bra ut.

2.3 Problem 3 - Konfidensintervall för Rayleighfördelning

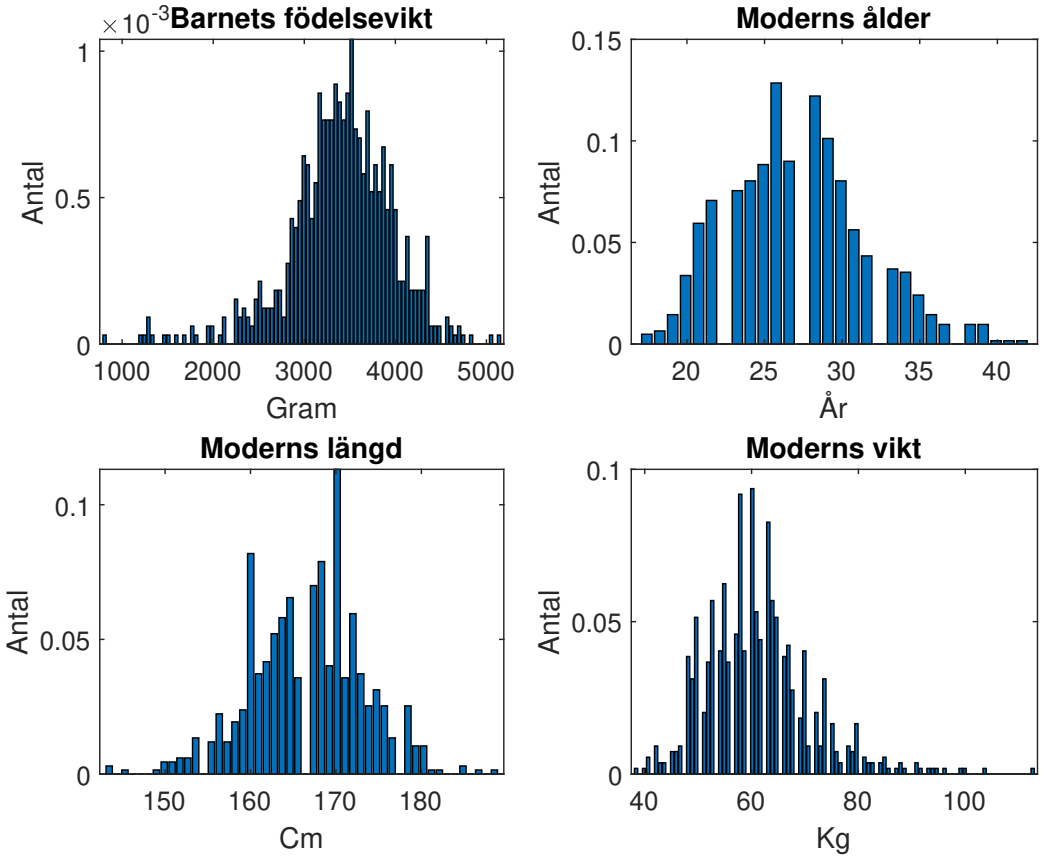


Figur 10: Simulering av skattning och gränser till konfidensintervall.

För att skatta parametern σ i datat använde vi oss av vår tidigare ML-skattning. Parametrarna som användes för skattningen är $n = 10000$, $\sigma = 1$ och $\alpha = 0.05$. Således är konfidensgraden för intervallet 95%. Vid exekvering av koden fick vi följande resutlat för våra parametrar $my\_est = 0.9882$, $lower\_bound = 0.9686$ och $upper\_bound = 1.0078$. Vi ser att vår skattning ligger inom vårt konfidensintervall. Fördelningen ser ut att passa bra med skattningen.

2.4 Problem 4 - Jämförelse av fördelningar hos olika populationer

2.4.1 Problem 4.1 - Generera histogram



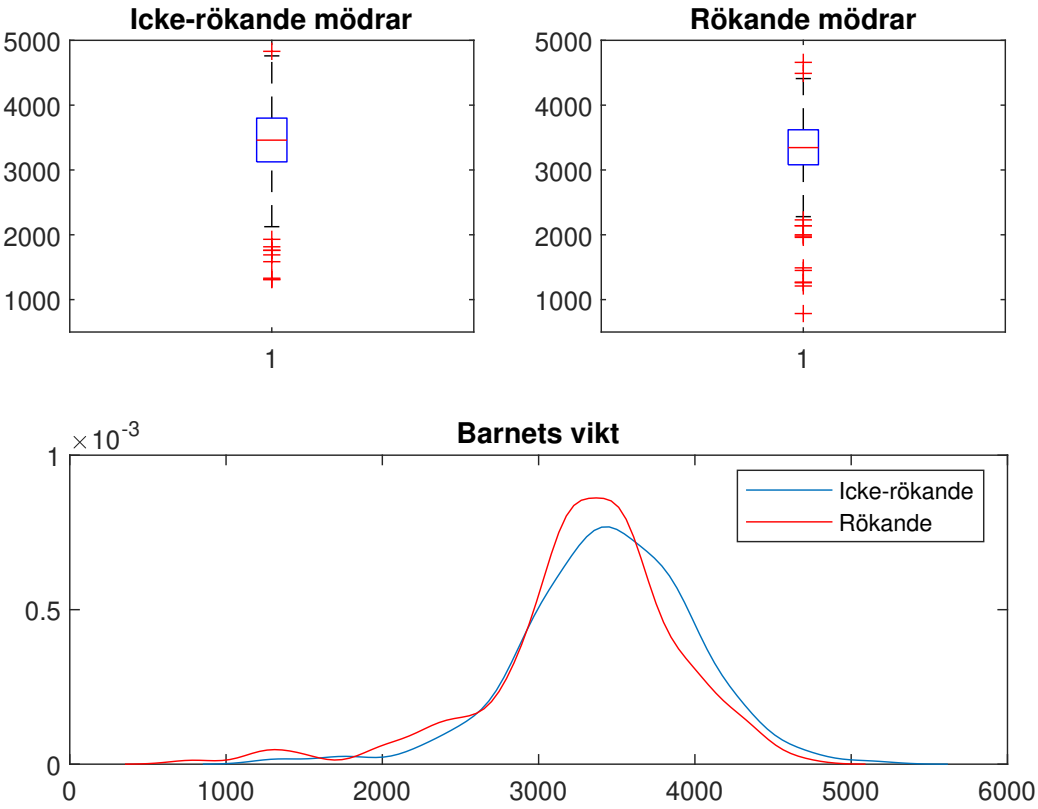
Figur 11: Olika histogram för förändringar hos moderns, vikt, ålder, längd och barnets födelsevikt.

Histogrammet för barnets vikt genereras med följande kod:

```
% Ladda all data från birth.dat
data = load("birth.dat");

birth_weight = data(:, 3);
subplot(2, 2, 1);
hist_density(birth_weight, 100);
title('Barnets födelsevikt');
xlabel('Gram');
ylabel('Antal');
```

Vikten för barn kan hittas i den tredje kolumnen i filen birth.dat.

2.4.2 Problem 4.2 - Rökande & icke-rökande mödrar



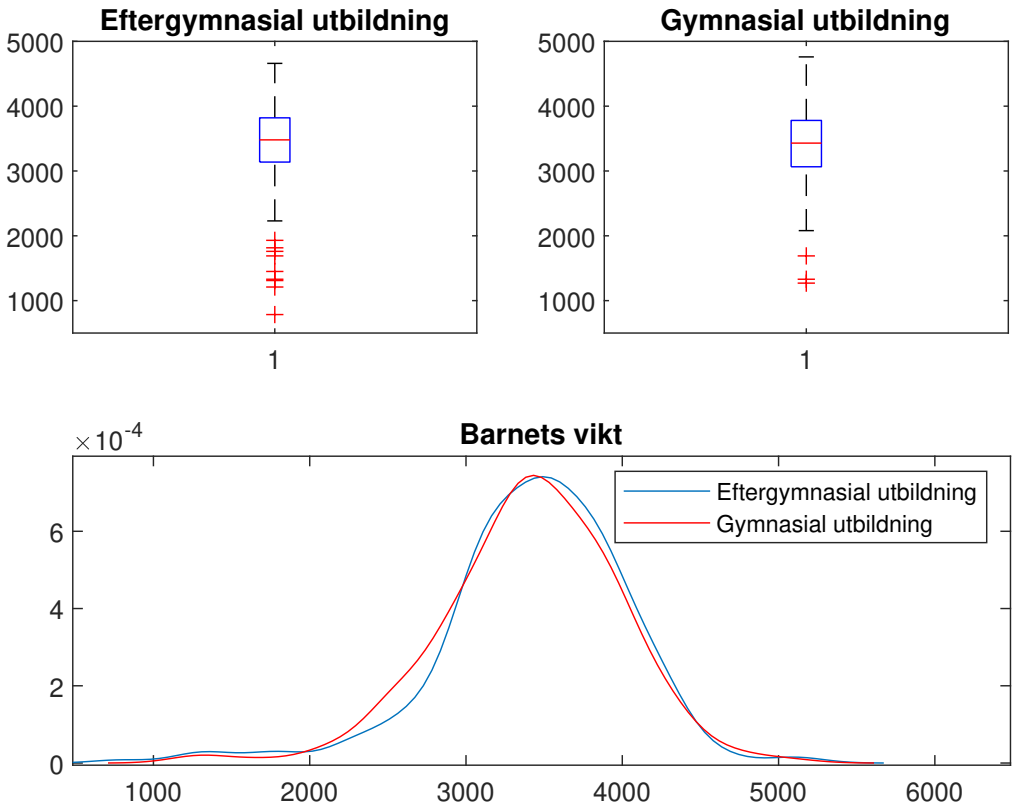
Figur 12: Risk för låg födelsevikt beroende på om modern röker eller ej.

De två låddiagrammen visar antalet mödrar som röker eller ej. Den blåa lådan representerar 50% av av all data. Övre strecket är således den tredje kvantilen (Q_3) och det undre är den första kvantilen (Q_1). Det röda strecket som finns i båda lådorna är medianen. Strecken som går ut från lådan kallas för *"whiskers"* all data som ligger som avviker från resterande data, detta kan ses som extremfall.

Den undre grafen presenterar barnets vikt beroende på om modern röker eller ej. Slutsatsen man kan dra utifrån diagrammet är att barn vars mor röker under graviditeten har en högre risk för låg födelsevikt.

2.4.3 Problem 4.2 - Utbildning

Vi skall nu studera om moderns utbildning är en riskfaktor för barnets födelsevikt. Vi väljer att studera mödrar med 10-12 års studier (gymnasial utbildning) och de med eftergymnasial utbildning.

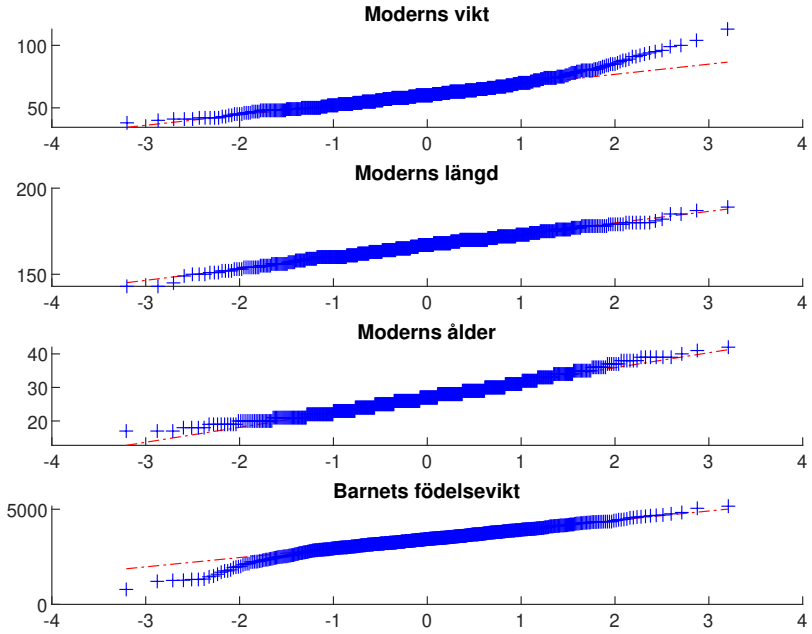


Figur 13: Risk för låg födelsevikt beroende på moderns utbildning.

Vi ser att det finns en tydlig risk för barnets födelsevikt ner till 2000g beroende på om modern endast har studerat 10-12 år.

2.5 Problem 5 - Test av normalitet

2.5.1 Problem 5.1 - Normalfördelning



Figur 14: Jämförelse med qqplot för olika variabler, med $\alpha = 0.05$.

Figuren visar att moderns vikt, längd, ålder och barnets födelsevikt är någorlunda normalfördelade. Barnets födelsevikt och moderns ålder avviker dock en del i början. Medan moderns vikt avviker i slutet. Den centrala-gränsvärdessatsen säger att om vi har oändligt många variabler kommer de vara normalfördelade. Detta är ett tydligt exempel, vi har många variabler men inte oändligt många.

2.5.2 Problem 5.2 - Statistiska test för normalitet

Detta problem behandlar hypotetsprövning. Vår nollhypotes är att moderns, längd, ålder, vikt och barnets födelsevikt är normalfördelade på signifikansnivån 5%. För att utföra hypotetsprövningen kommer vi att använda oss av Jarque-Beras test av normalitet och funktionen `jbtest` i Matlab.

Jbtest kommer att returnera antingen 1 eller 0, men man kan även returnera exempelvis P-värde om man så vill. Således finns det fler sätt att angripa problemet på. Vi kommer att hålla det simpelt och bara studera om vi får 1 eller 0. Om funktionen returnerar 1 så skall H_0 förkastas. Om funktionen returnerar 0 så skall H_0 **inte** förkastas.

För att utföra testet för normalitet använder vi följande kod:

```
1    significance_level = 0.05;
2    mother_length_nd = jbtest(mother_length, significance_level);
3    mother_age_nd = jbtest(mother_age, significance_level);
4    birth_weight_nd = jbtest(mother_age, significance_level);
5    mother_weight_nd = jbtest(mother_weight, significance_level);
6
```

```
    fprintf('Moderns vikt är normalfördelad: %s\n',
            tfToString(mother_length_nd == 0));
    fprintf('Moderns ålder är normalfördelad: %s\n',
            tfToString(mother_age_nd == 0));
    fprintf('Moderns vikt är normalfördelad: %s\n',
            tfToString(mother_weight_nd == 0));
    fprintf('Barnets födelsevikt är normalfördelad: %s\n',
            tfToString(birth_weight_nd == 0));
```

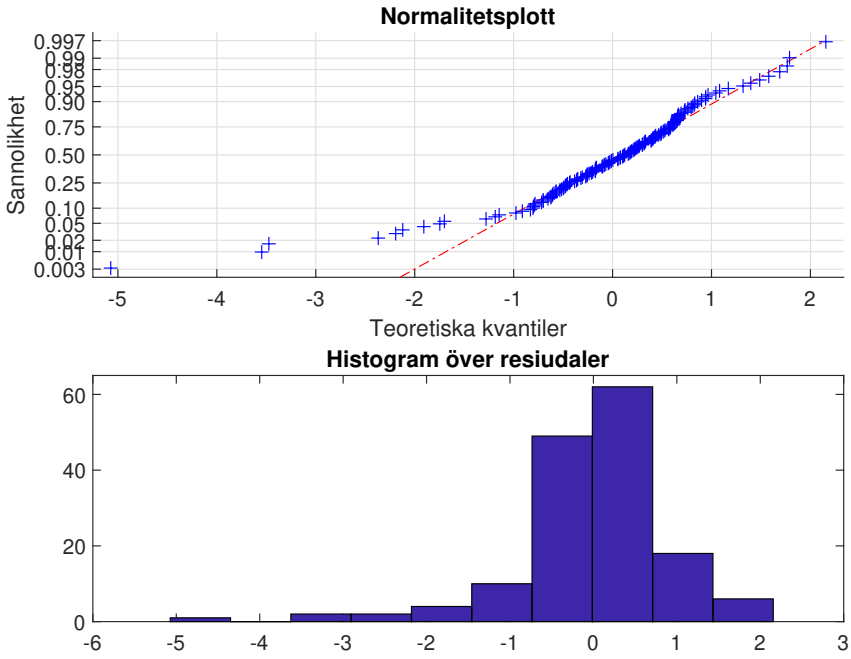
När vi exekverar koden får vi följande resultat:

```
    Moderns längd är normalfördelad: sant.
    Moderns ålder är normalfördelad: falskt.
    Moderns vikt är normalfördelad: falskt.
    Barnets födelsevikt är normalfördelad: falskt.
```

Slutsatsen är att endast moderns längd är normalfördelad, medan moderns ålder, vikt och barnets födelsevikt inte är det. Detta kunde vi inte se med graferna i Figur 14. Jarque-Beras test av normaltitet ger ett bättre svar på om datan är normalfördelad.

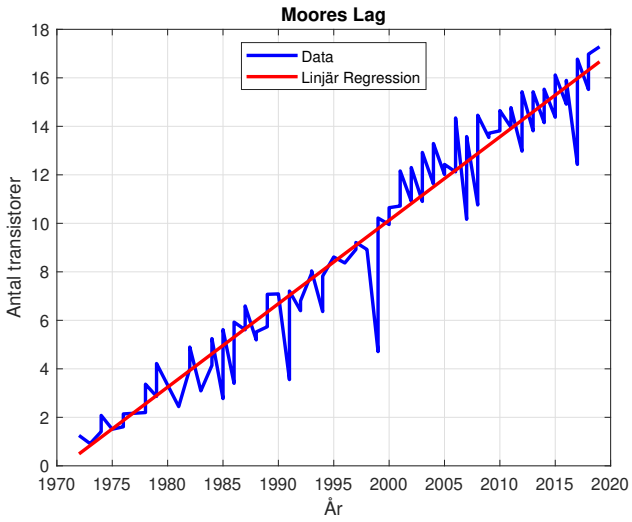
2.6 Problem 6 - Enkel linjär regression

2.6.1 Problem 6.1 - Regression & residualanalys



Figur 15: Normalitetsplott och histogram för residualer.

En resiudual är skillnaden mellan praktiskt värde och teoretiskt värde. Residualerna ser ut att komma från en normalfördelning. De har dock en avvikelse i början av normalfördelningskurvan. Analysen av residualerna tyder på att regressionen är någorlunda lämplig, ty de ser ut att vara normalfördelade.



Figur 16: Plott av den linjära regressionen.

Vi använder följande kod för regressionen:

```
trans_per_area = data(:, 1);
year = data(:, 2);

w = log(year);

X = [ones(length(trans_per_area), 1), trans_per_area];

beta_hat = regress(w, X);
```

2.6.2 Problem 6.2 - Storheten R^2

Storheten R^2 är ett mått på hur stor andel av variationen i datan som förklaras av modellen. Storheten får med formeln

$$R^2 = 1 - \frac{summa\ av\ kvadrerad\ regressionsvariation\ (SSR)}{total\ summa\ av\ kvadrater\ (SST)}$$

$$R^2 = 1 - \frac{SSR}{SST}$$

Vi använder följande kod för att beräkna R^2:

```
[beta_hat, ~, ~, ~, stats] = regress(w, X);

R_squared = stats(1);
```

Vårt resultat är $R^2 = 0.9586$.

2.6.3 Problem 6.3 - Förutsägelse för transistorer år 2025

Vi annvänder oss av följande kod för att förutse antal transistorer år 2025:

```
year_2025 = 2025;
X_2025 = [1, year_2025];
log_prediction_2025 = X_2025 * beta_hat;
prediction_2025 = exp(log_prediction_2025);
```

Vi skapar ett x-värde för år 2025. Sedan gör vi en förutsägelse för det logaritmiska värdet av det specifika året, genom att multiplicera matrisen med $\hat{\beta}$. Slutligen omvandlar vi det logaritmiska värdet till den ursprungliga skalan.

Antal transistorer:

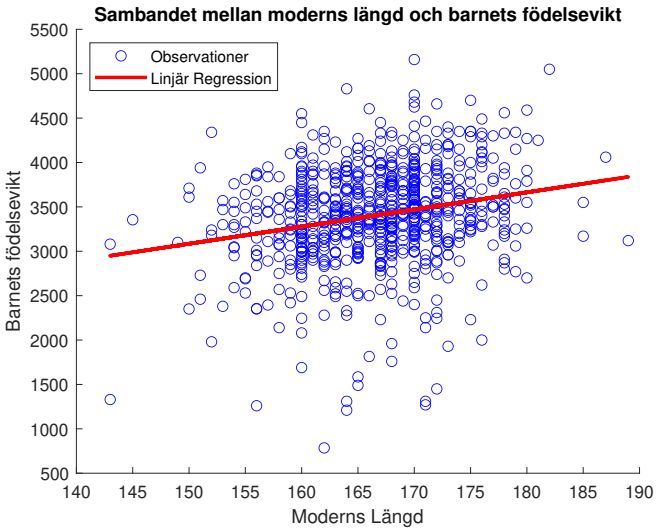
- **år 2021**: 34 337 892 st.
- **år 2025**: 135 986 733 st.

Våra beräkningar stämmer överens med vad som kan hittas på internet. Således är vår förutsägelse adekvat.

2.7 Problem 7 - Multipel linjär regression

2.7.1 Problem 7.1 - Enkel linjär regression med moderns vikt

Vi skall nu sätta upp en enkel linjär regerssionsmodell för hur barnets födelsevikt beror på moderns längd. Grafen nedan presenterar sambandet.



Figur 17: Enkel linjär regression.

Denna graf är generarad med följande kod:

```
X = [ones(length(mother_length), 1), mother_length];

beta_hat = regress(birth_weight, X);

% Plotta sambandet mellan moderns längd och födelsevikten
scatter(mother_length, birth_weight, 'o', 'filled',
        'DisplayName', 'Observationer');
hold on;

% Lägg till linjär regressionslinje
plot(mother_length, X * beta_hat, 'r',
    'LineWidth', 2, 'DisplayName', 'Linjär Regression');
```

2.7.2 Problem 7.2 - Multipel linjär regression

Vi skall nu konstruera en multipel linjär regressionsmodell där förklaringsvariablerna är moderns vikt, moderns rökvanor och utbildning. Modellen skapas med följande kod:

```
mw = data(:, 15);
ms = data(:,20) == 3;
mhs = data(:,24) == 2;
birth_weight = data(:,3);

X = [ones(size(mother_weight)),mw,ms,mhs];
[beta_hat, confidence_interval, res] = regress(birth_weight, X, 0.05);
```

Konfidensintervallen kan direkt erhållas med hjälp av funktionen `regress`. I detta fall har vi ett konfidensintervall med konfidensgraden 95%, eftersom $\alpha = 0.05$.

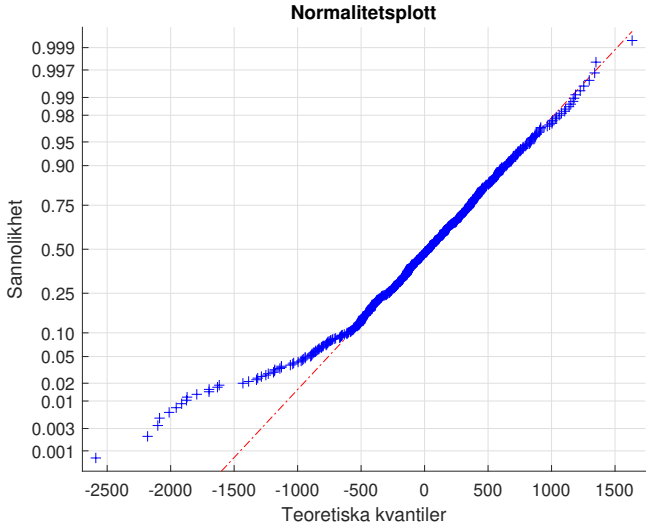
Grafen går dessvärre inte att plotta, eftersom den är fyrdimensionell. Regressionen beror av tre variabler.

När vi skattar koefficenterna och konfidensintervallen får vi följande resultat:

```
Uppskattade koefficienter:
1.0e+03 *

 2.7681
 0.0113
-0.1622
-0.0259

Konfidensintervall för koefficienter:
1.0e+03 *

 2.5102    3.0260
 0.0072    0.0154
-0.2540   -0.0705
-0.1176    0.0657
```

Slutsatsen som vi kan dra med denna uppskattning är att moderns vikt påverkar barnets födelsevikt positivt. Mödrar som röker och har endast en gymnasial utbildning påverkar barnets vikt negativt. Vi ser även att alla skattningar ligger innanför vårt konfidensintervall, med $\alpha = 0.05$.

Nedan plottar vi residualerna från regerssionen.



Figur 18: Normalitetsplott av residualerna.

Slutsatsen som vi kan dra utifrån grafen är residualerna ser ut att komma från en normalfördelning men att de har en avvikelse i början. Det kan exempelvis finnas ett systematiskt beteende i början eller att det finns mycket data som ligger utanför. Det är viktigt att avgöra varför dessa avvikelser finns. Det kan exempelvis vara mätfel eller oregelbundet beteende. I vårt fall är det avvikande, eftersom barnets vikt inte är linjärt beroende av alla variabler.

Att residualerna är någorlunda normalfördelade tyder på att vår regression är relativt adekvat för detta fall.