Laboration 3: Statistisk prediktion av global temperatur

Sebastijan Babic

2025-01-05

Contents

Sammanfattning

I denna laboration har vi analyserat globala temperaturdata för att identifiera trender och göra prediktioner. Vi har använt linjär regression för att undersöka sambandet mellan temperatur och årtal, och jämfört olika modeller för att bedöma deras lämplighet för prediktioner. Resultaten visar att modeller baserade på nyare data ger mer exakta prediktioner för moderna perioder.

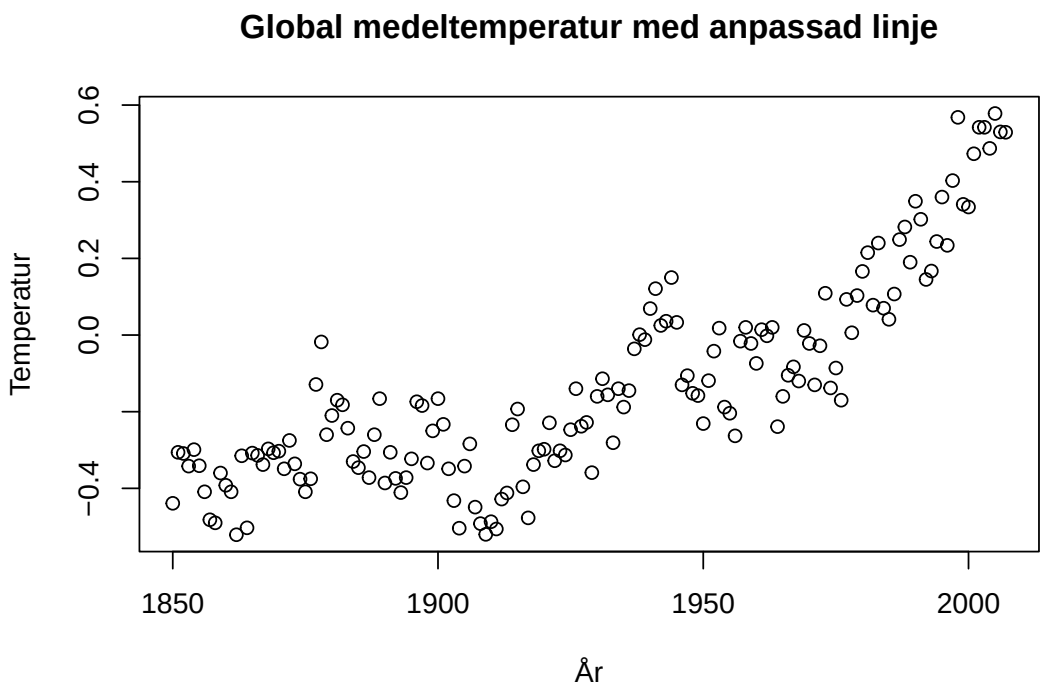
Vi har kommit framm till att en modell baserad på data från 1970–2007 ger bättre prediktioner för perioden 2008–2022 jämfört med en modell baserad på data från 1850–2007. Detta beror på att den senare modellen inkluderar en längre tidsperiod med större variation i temperaturmönster, vilket leder till högre felvärden när den används på modern data. För att förbättra prediktionerna kan vi överväga att använda mer komplexa modeller som tar hänsyn till icke-linjära mönster och fler förklarande variabler.

Uppgift 1 - Jordens medeltemperatur 1850–2007

```r
df <- read.csv("temperatur.csv", header = TRUE)
```

```r
temperature <- df$temperatur
age <- df$år

plot(age, temperature, xlab = "År", ylab = "Temperatur", main = "Global medeltemperatur med anpas
```



Diagrammet visar en ökning i den globala medeltemperaturen över tid. Det ser ut att det finns en positiv trend, men ett linjärt samband är inte direkt uppenbart eftersom temperaturen verkar accelerera mot slutet av perioden. Detta kan tyda på att sambandet snarare är icke-linjärt (t.ex. exponentiellt).

För att vi ska kunna göra en linjär regression så har vi kraven:

Krav för linjär regression:

1. **Linjäritet**: Sambandet mellan variablerna bör vara linjärt.
2. Residualerna bör ha **konstant varians**.
3. **Oberoende observationer**: Inga observationer bör påverka varandra.

4. **Normalfördelade residualer**: Residualerna bör följa en normalfördelning.

Vi genomför analys för att se om det finns något samband mellan temperatur och år och sedan ser om det är rimligt med ett linjär regression genom att utföra residualanalys och analys via normalfördelningsplot.

```
old_par <- par(mfrow = c(1, 3))
modell <- lm(temperature ~ 1 + age, data = df) # vill ha intercept så + 1

# scatterplot med linje
plot(age, temperature, xlab = "År", ylab = "Temperatur", main = "Global medeltemperatur")
abline(modell, col = "red")

# residual plot
residual <- modell$residuals
plot(age, residual, xlab = "År", ylab = "Residualer", main = "Residualplot")
abline(a = 0, b = 0, lty = "dotted")

# normalfördelningsplot
qqnorm(residual, main = "Normalfördelningsplot residualer")
qqline(residual)
```

```
par(old_par)
```

Vi ser redan i första plotten till vänster att vi verkar inte följa en linjär trend pga. den exponentialla ökningen mot slutet av plotten. Vi ser även i residualplotten att det inte finns någon konstant värde pga. den förändringen i avståndet från 0-linjen som inte är konstant, mao. spridningen verkar förändras i olika delar av plotten och det verkar bildas en sorts "mönster" i spridningen, vi bildar nästan ett W. Residualerna verkar följa en normalfördelning förutom vid dem extrema värden i svansarna.

Vi har därmed inte uppfyllt kraven på linjär regression och vi kan inte dra några slutsatser från denna modell.

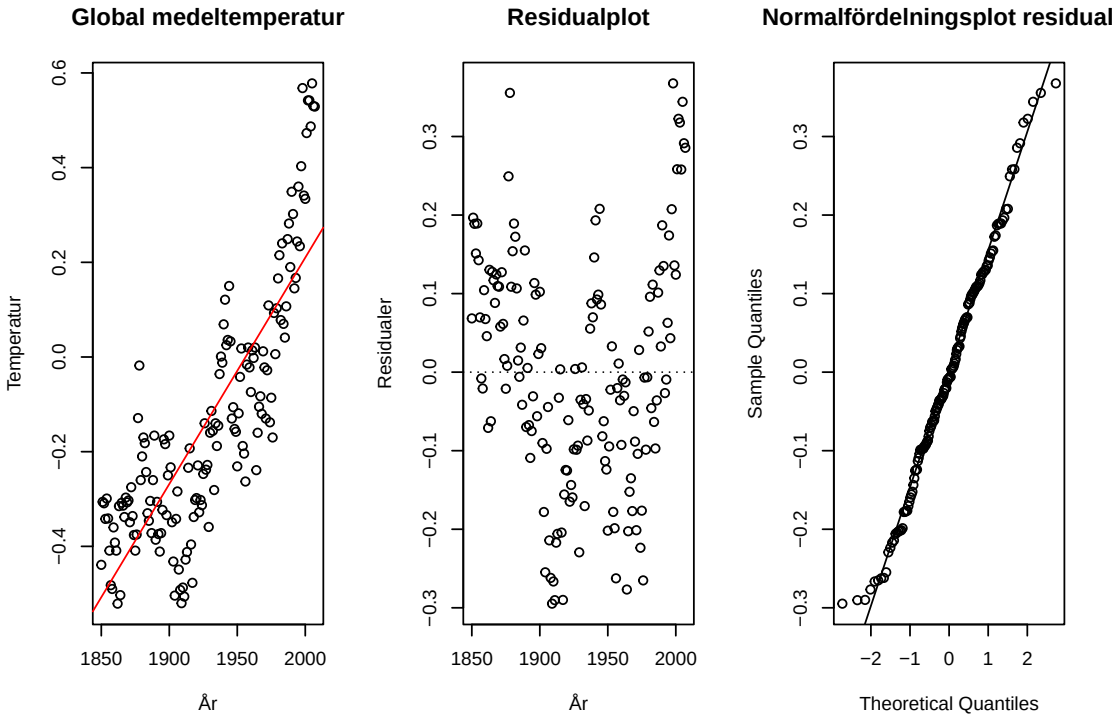


Figure 1: Linjär regression, residualanalys och QQ-plot analys för global temperaturdata

Uppgift 2 - Jordens medeltemperatur under tre perioder

```
# delar in i tre årsperioder
df1 <- subset(df, år >= 1880 & år <= 1929)
df2 <- subset(df, år >= 1930 & år <= 1969)
df3 <- subset(df, år >= 1970 & år <= 2007)

# utför linjär regression för varje årsperiod för att se om kraven uppfylls bättre
modell1 <- lm(temperatur ~ år, data = df1)
modell2 <- lm(temperatur ~ år, data = df2)
modell3 <- lm(temperatur ~ år, data = df3)

old_par <- par(mfrow = c(1, 2))
# summering av modellerna
residual1 <- modell1$residuals
plot(df1$år, residual1, xlab = "År, 1880-1929", ylab ="Residualer", main = "Residualplot")
abline(a = 0, b = 0, lty = "dotted")
qqnorm(residual1)
qqline(residual1)
```

```
old_par <- par(mfrow = c(1, 2))
residual2 <- modell2$residuals
plot(df2$år, residual2, xlab = "År, 1930-1969", ylab ="Residualer", main = "Residualplot")
abline(a = 0, b = 0, lty = "dotted")
qqnorm(residual2)
qqline(residual2)
```

```
old_par <- par(mfrow = c(1, 2))
residual3 <- modell3$residuals
plot(df3$år, residual3, xlab = "År, 1979-2007", ylab ="Residualer", main = "Residualplot")
abline(a = 0, b = 0, lty = "dotted")
qqnorm(residual3)
qqline(residual3)
```

Vi verkar komma närmare att uppfylla kraven för linjär regression för varje årsperiod. Vi ser att residualerna verkar ha en konstant varians pga. den slumpvisa spridningen och en någorlunda normalfördelning på QQ-plotten.

Där vi använder `summary` för att få fram interceptet och lutningskoefficienten för varje modell.

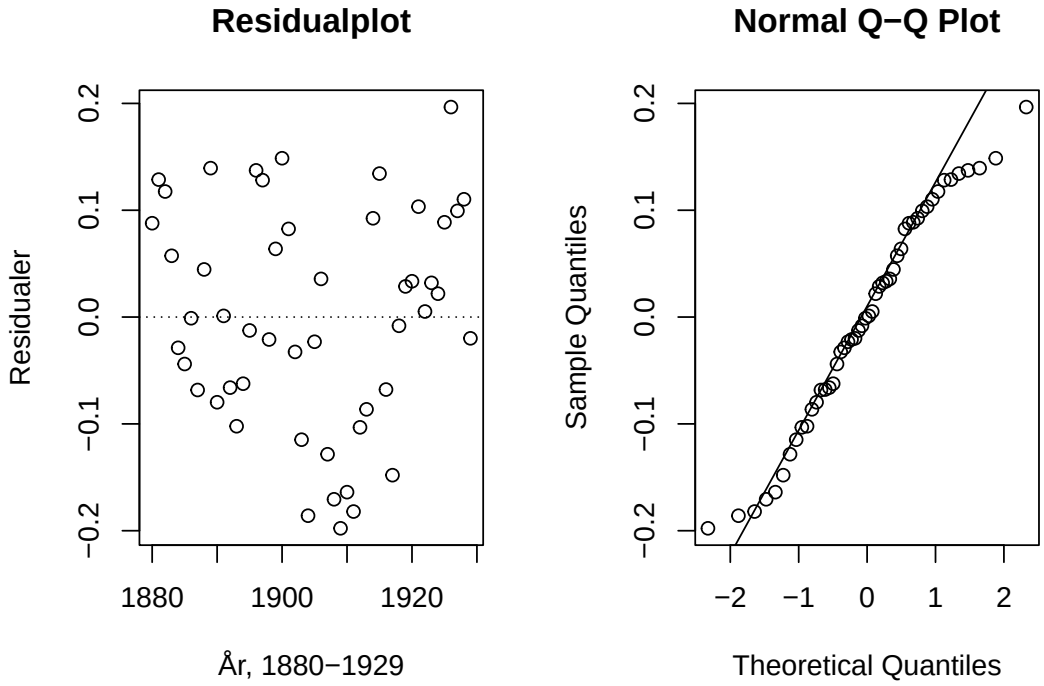


Figure 2: Residualplot och QQ-plot för olika årsperioder

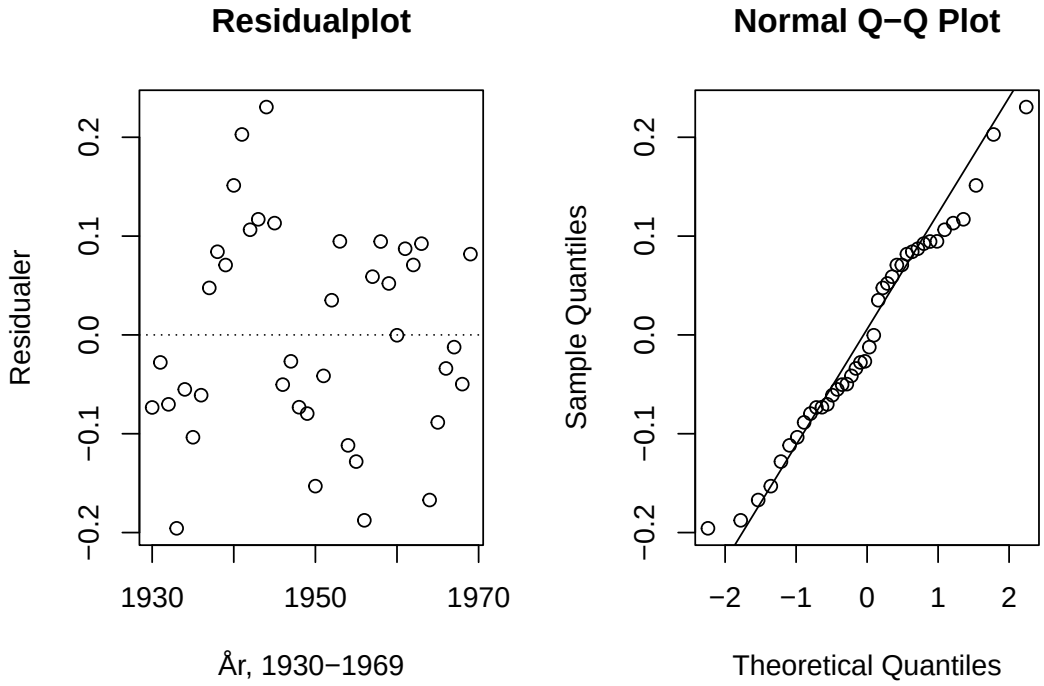


Figure 3: Residualplot och QQ-plot för olika årsperioder

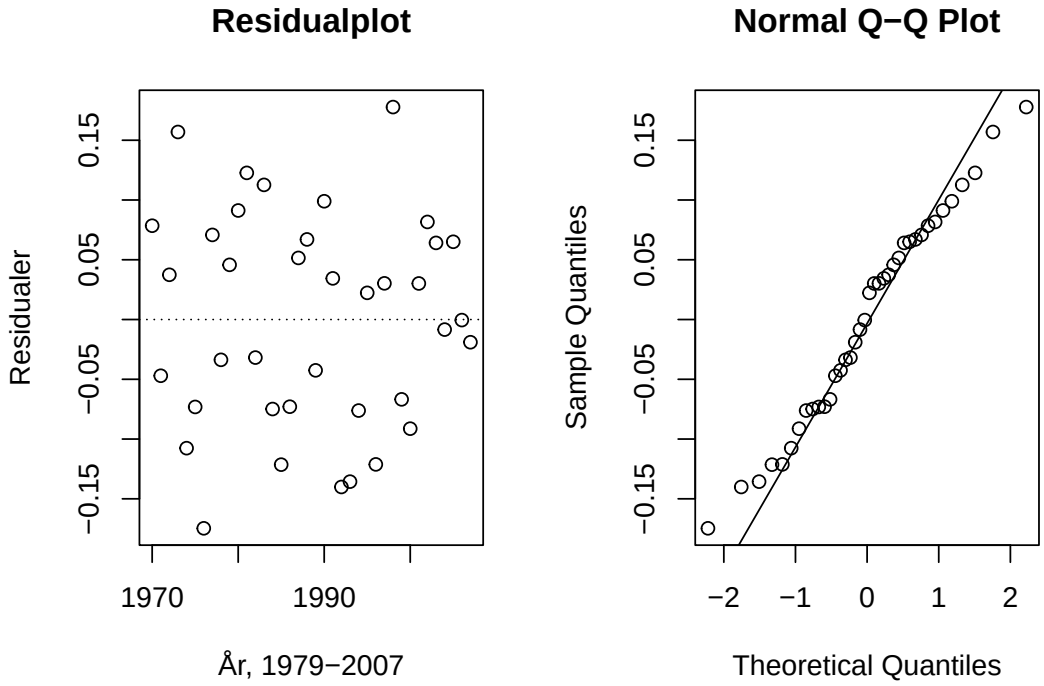


Figure 4: Residualplot och QQ-plot för olika årsperioder

```r
# Extraherar koefficienter
coef1 <- coef(modell1)
coef2 <- coef(modell2)
coef3 <- coef(modell3)

# Extraherar intercept och lutningskoefficienter
intercept1 <- coef1[1]
slope1 <- coef1[2]

intercept2 <- coef2[1]
slope2 <- coef2[2]

intercept3 <- coef3[1]
slope3 <- coef3[2]

# Skapar en data.frame för tabellen
table_data <- data.frame(
  Modell = c("Modell 1 (1880-1929)", "Modell 2 (1930-1969)", "Modell 3 (1970-2007)"),
  Intercept = c(intercept1, intercept2, intercept3),
  Lutningskoefficient = c(slope1, slope2, slope3)
)

# Visar tabellen med knitr::kable
knitr::kable(
  table_data,
  caption = "Sammanställning av intercept och lutningskoefficient för tre modeller.",
  col.names = c("Modell", "Intercept", "Lutningskoefficient"),
  digits = 5  # Begränsar decimaler till 3
)
```

Table 1: Sammanställning av intercept och lutningskoefficient för tre modeller.

Modell	Intercept	Lutningskoefficient
Modell 1 (1880–1929)	1.28498	-0.00084
Modell 2 (1930–1969)	-0.91997	0.00043
Modell 3 (1970–2007)	-34.62775	0.01753

Vi vill nu göra en hypotesprövning för att se om det finns en trend mot varmare klimat ($\beta > 0$) med en ensidig hypotesprövning för perioden 1970-2007 (modell 3) just pga. vi undersöker om

det förekommer varmare klimat. Vi använder oss av ett t-test för β för att se om det finns en signifikant trend.

Vi har alltså nollhypotesen $H_0 : \beta \leq 0$ och alternativhypotesen $H_1 : \beta > 0$. Vi använder oss av `summary()` för att få fram p-värdet.

```
p_value <- summary(modell3)$coefficients[2, 4] / 2 # dela på två eftersom summary ger tvåsidig te
if (p_value < 0.05) {
  print("Vi förkastar nollhypotesen. Direktören har fel.")
} else {
  print("Vi kan inte förkasta nollhypotesen. Direktören kan ha rätt.")
}
```

```
## [1] "Vi förkastar nollhypotesen. Direktören har fel."
```

Uppgift 3 - Prediktion av global temperatur

Upgift 3.1 - Modellval

```r
df_test <- read.csv("temperatur_test.csv", header = TRUE)

# Modell 1: 1850-2007
modell_1850_2007 <- lm(temperatur ~ år, data = subset(df, år <= 2007))
pred_1850_2007 <- predict(modell_1850_2007, newdata = df_test)
rmse_1850_2007 <- sqrt(mean((df_test$temperatur - pred_1850_2007)^2))

# Modell 2: 1970-2007
modell_1970_2007 <- lm(temperatur ~ år, data = subset(df, år >= 1970 & år <= 2007))
pred_1970_2007 <- predict(modell_1970_2007, newdata = df_test)
rmse_1970_2007 <- sqrt(mean((df_test$temperatur - pred_1970_2007)^2))

# RMSE för 1970-2007 (träningsdata)
train_1970_2007 <- subset(df, år >= 1970 & år <= 2007)
pred_train_1850_2007 <- predict(modell_1850_2007, newdata = train_1970_2007)
rmse_train_1850_2007 <- sqrt(mean((train_1970_2007$temperatur - pred_train_1850_2007)^2))

pred_train_1970_2007 <- predict(modell_1970_2007, newdata = train_1970_2007)
rmse_train_1970_2007 <- sqrt(mean((train_1970_2007$temperatur - pred_train_1970_2007)^2))

library(knitr)

# skapa RMSE-tabell
rmse_table <- data.frame(
  Modell = c("1850-2007", "1970-2007"),
  `2008-2022 (Testdata)` = c(round(rmse_1850_2007, 3), round(rmse_1970_2007, 3)),
  `1970-2007 (Träningsdata)` = c(round(rmse_train_1850_2007, 3), round(rmse_train_1970_2007, 3))
)

# skriv ut tabellen med knitr::kable
kable(
  rmse_table,
  col.names = c("Modell", "Testdata (2008-2022)", "Träningsdata (1970-2007)"),
  align = c('l', 'c', 'c'),
  booktabs = TRUE
)
```

Modell	Testdata (2008–2022)	Träningsdata (1970–2007)
1850–2007	0.538	0.179
1970–2007	0.151	0.088

3.1 - Slutsats

Modellen 1850–2007 Modellen baserad på 1850-2007 har ett RMSE på 0.538 vilket är relativt högt. Det kan bero på att modellen inkluderar från en mycket lång period med stor variation i temperaturmönster som kan leda till ett mindre anpassad modell. Modellen baserad på 1970-2007 har ett lägre RMSE på 0.151 vilket kan bero på att modellen är mer anpassad till den senaste perioden med mindre variation i temperaturmönster och är därför mer lämplig att använda i att prediktera framtida temperaturer.

Modellen baserad på 1970-2007 har ett lägre RMSE, 0.179 vilket är lägre än 0.538 såsom innan men fortfarande ganska högt jämfört med modellen 1970-2007 med ett RMSE på 0.088 som är rimligt eftersom den modellen använder mer "modern" data och kan därför tillämpas bättre på framtida data.

Skillnaden i felvärden beror på vilken data modellerna anpassats till. Modellen 1850–2007 använder en mycket bred tidsperiod, vilket gör att den fångar både långsiktiga trender och variationer som inte är relevanta för de senaste decennierna. Detta resulterar i högre fel när den utvärderas på moderna data (2008–2022). Modellen 1970–2007 fokuserar endast på moderna data och därmed bättre fångar temperaturökningen under denna period. Den är därför mer exakt för att prediktera moderna temperaturer.

Några fördelad med att använda en modell med mer data i modellanpassningen är att den kan fånga långsiktiga trender. Det kan till exempel vara cykliska mönster som inte är uppenbara i kortare tidsperioder. Nackdelen är att modellen kan bli överanpassad och få högre felvärden när den används på modern data. Med andra ord så kan modellen vara felaktig på grund av förändrare miljöförhållande exempelvis.

Modellen 1970–2007 Modellen baserad på 1970–2007 har ett mycket lågt RMSE på 0.088, vilket är förväntat eftersom modellen är anpassad till exakt samma data. Detta visar att den har en mycket god passform för perioden den är tränad på. Modellen 1850–2007 har ett något högre RMSE på 0.179, vilket beror på att den inte är lika specifikt anpassad till perioden 1970–2007.

Resterande av slutsatsen är snarlik som för modellen 1850-2007 och därför hänvisas det till ovan.

Uppgift 3.2 - Prediktion och prediktionsintervall

```r
# prediktionsintervallen
pred_intervals <- predict(modell_1970_2007, newdata = df_test, interval = "prediction")

plot(df_test$år, df_test$temperatur, pch = 16, col = "blue", xlab = "År", ylab = "Temperatur",
     main = "Prediktion och prediktionsintervall (2008-2022)")
lines(df_test$år, pred_intervals[, "fit"], col = "red", lwd = 2)
lines(df_test$år, pred_intervals[, "lwr"], col = "green", lty = 2)
lines(df_test$år, pred_intervals[, "upr"], col = "green", lty = 2)

legend("topleft", legend = c("Faktiska värden", "Prediktion", "Prediktionsintervall"),
       col = c("blue", "red", "green"), pch = c(16, NA, NA), lty = c(NA, 1, 2), bty = "n")
```

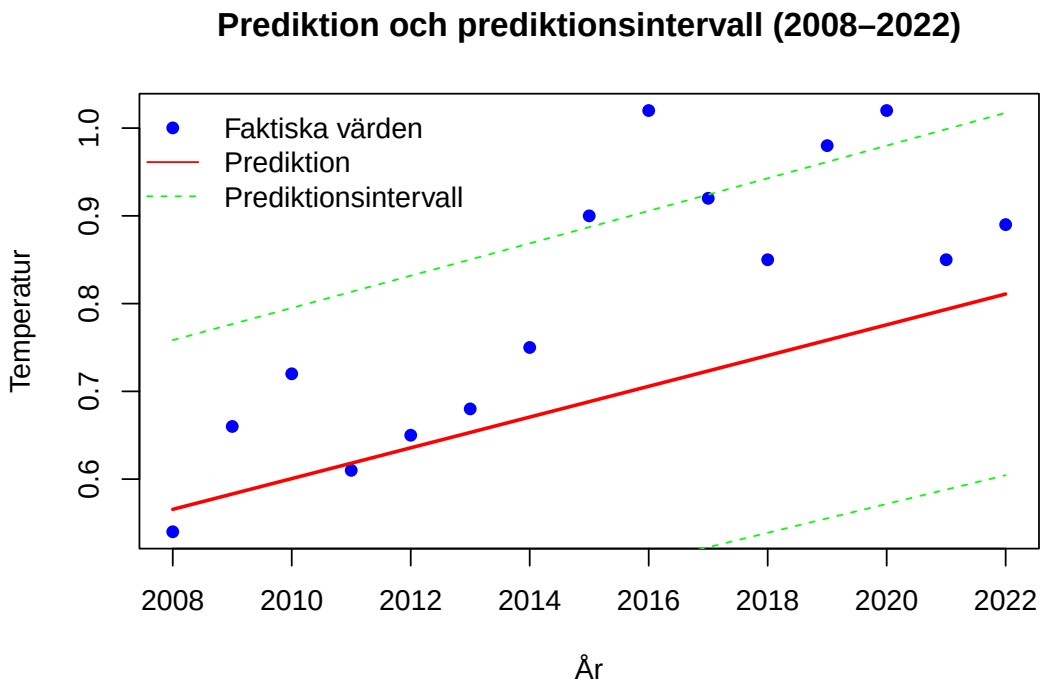


Figure 5: Prediktion och prediktionsintervall för global temperatur (2008–2022)

De faktiska temperaturvärdena verkar avvika från de predikterade värdena (röda linjen) för flera år, särskilt mot slutet av perioden (2008–2022). Detta antyder att en linjär modell inte fångar de icke-linjära mönstren i data tillräckligt väl. Prediktionsintervallet (gröna streckade

linjer) är brett, vilket visar att osäkerheten i förutsägelserna är stor. Detta minskar modellens användbarhet för exakta prediktioner.

För att eventuellt förbättra modellen kan vi dela upp datan i mindre tidsperioder och använda olika modeller för varje period för att bättre spegla förändringar i klimattrender över tid. Vi skulle även kunna införa flera förklarande variabler som skulle ge en mer komplex modell som bättre kan fånga upp de icke-linjära mönstren i data.

Några problem med att prediktera framtida temperatur på detta sätt kan vara:

1. Överförenkling, genom att endast använda år som prediktor ignoreras viktiga faktorer som påverkar temperaturer, exempelvis mänskliga utsläpp och naturliga variationer.
2. Osäkerhet i framtida händelser, framtida klimatpolitik, teknologiska framsteg och naturfenomen som vulkanutbrott är oförutsägbara och kan påverka temperaturen dramatiskt, något som Nassim Nicholas Taleb i sin bok “The Black Swan” tar upp.
3. Datakvalitet, i och med att vi har mät data från och med 1870 så kan det ha uppstått mätfel pga felaktiga mätinstrument.
4. Antagandet om konstant trend, linjär regression förutsätter en konstant ökning eller minskning över tid, vilket inte stämmer med de observerade icke-linjära mönstren i temperaturen.