Laboration 2: Statistisk hypotesprövning

Sebastian Babic

2024-12-31

Contents

Sammanfattning

För att göra linjär regression använder vi oss av funktionen `lm()` i R. Vi kan använda oss av `summary()` för att få en sammanfattning av modellen. För att göra en hypotesprövning använder vi oss av `anova()` och `summary()`. Vi kan också använda oss av `confint()` för att få konfidensintervall för parametrarna.

I `lm()` så skriver vi in vår modell som `lm(y ~ 1 + x, data = data)`. Där `y` är den beroende variabeln och `x` är den oberoende variabeln.

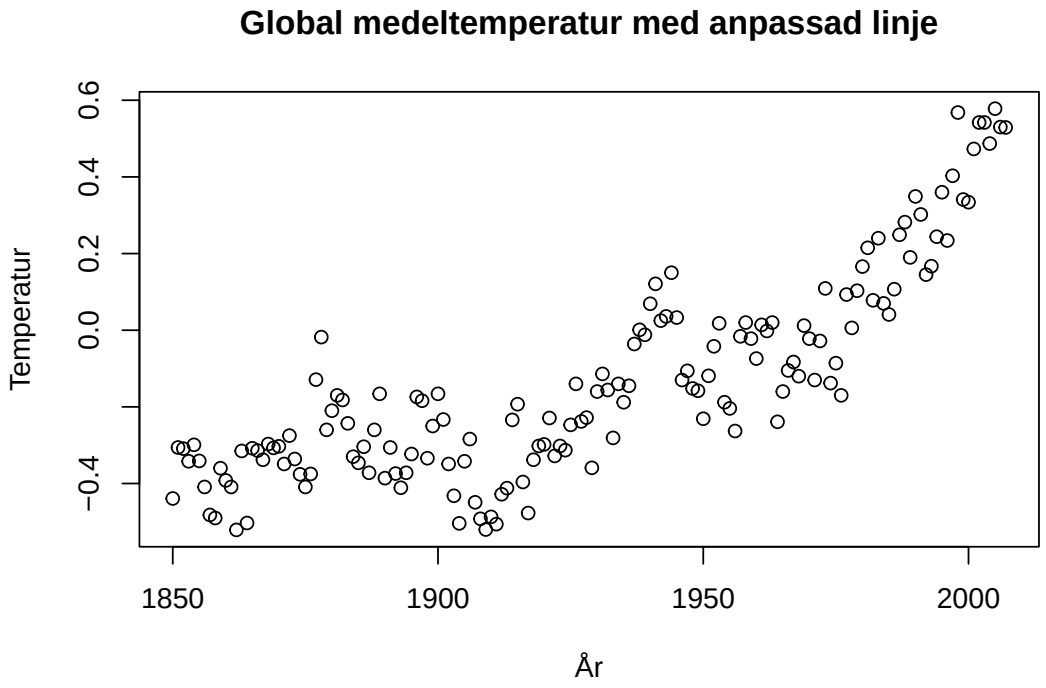
Använd `summary()` för att få fram värde.

För att få residueerna så kan vi använda oss av `residuals <- modell$residuals` och sedan plotta residuelaplotten med `plot(df$age, residuals)` ofta tsm med `abline(a = 0, b= 0, lty = "dotted"`.

Uppgift 1

```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:

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.

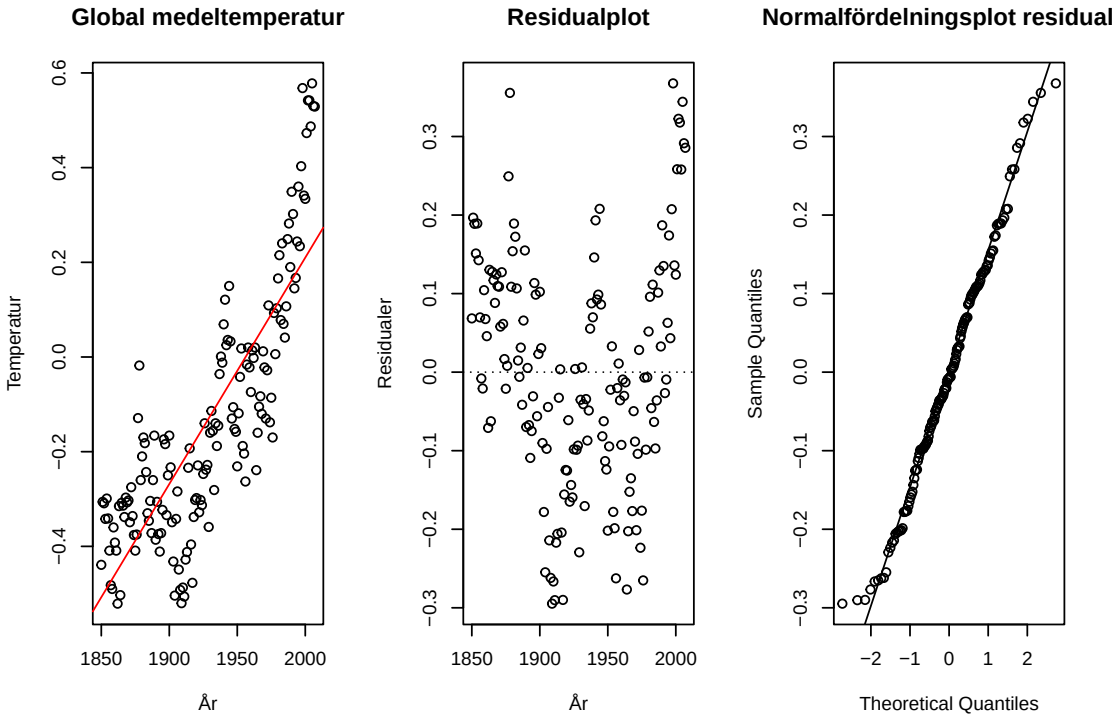
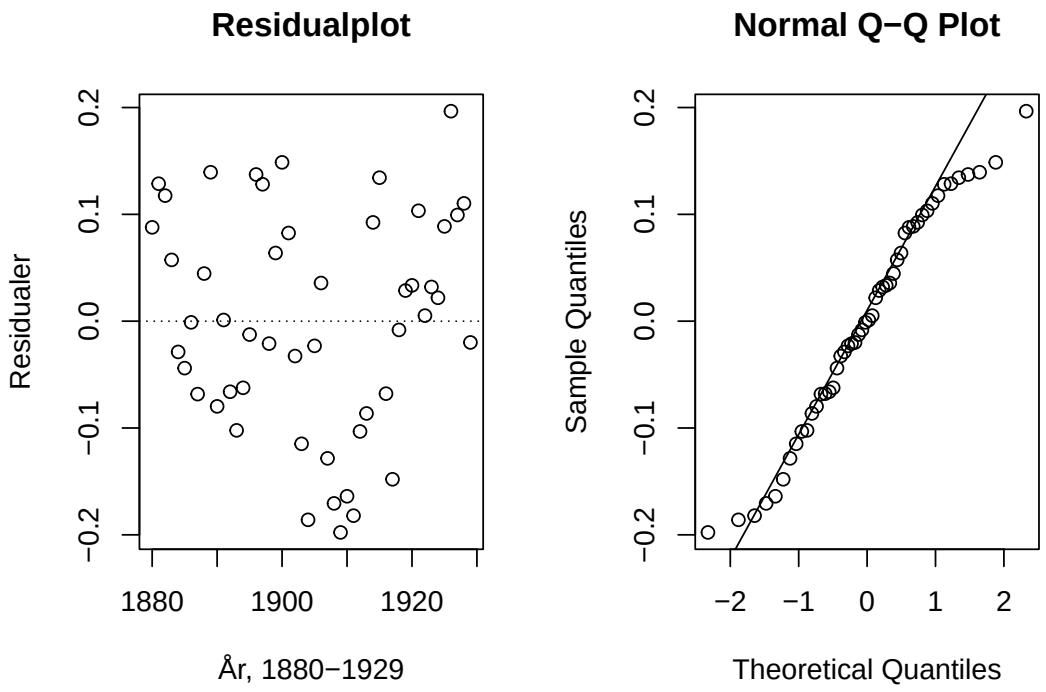


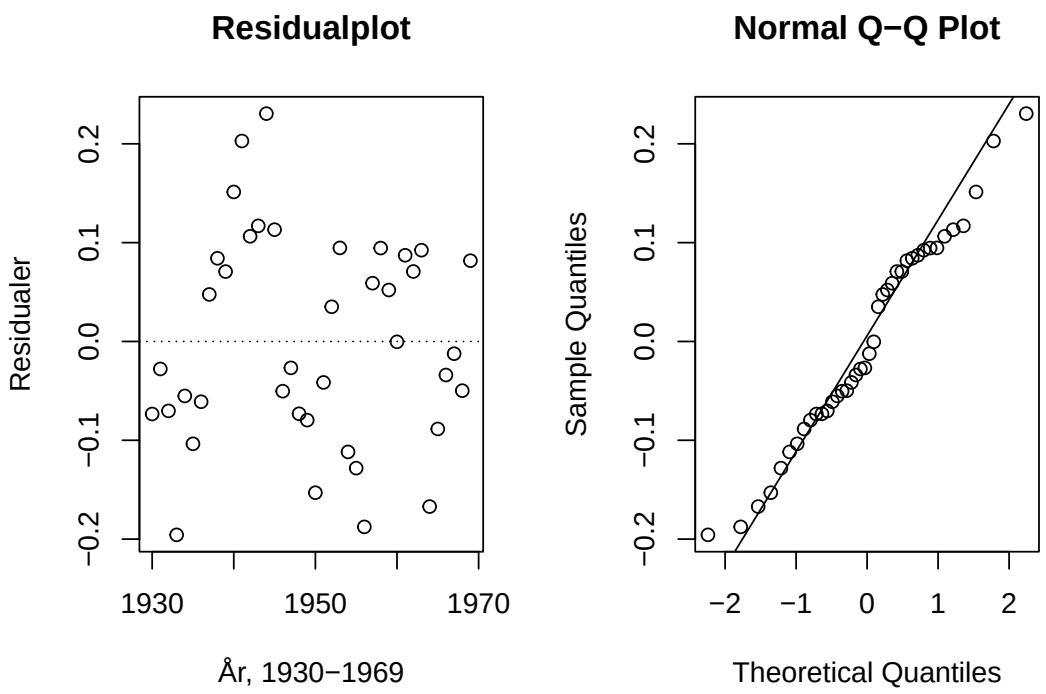
Figure 1: Linjär regression, residualanalys och QQ-plot analys för global temperaturdata

Uppgift 2

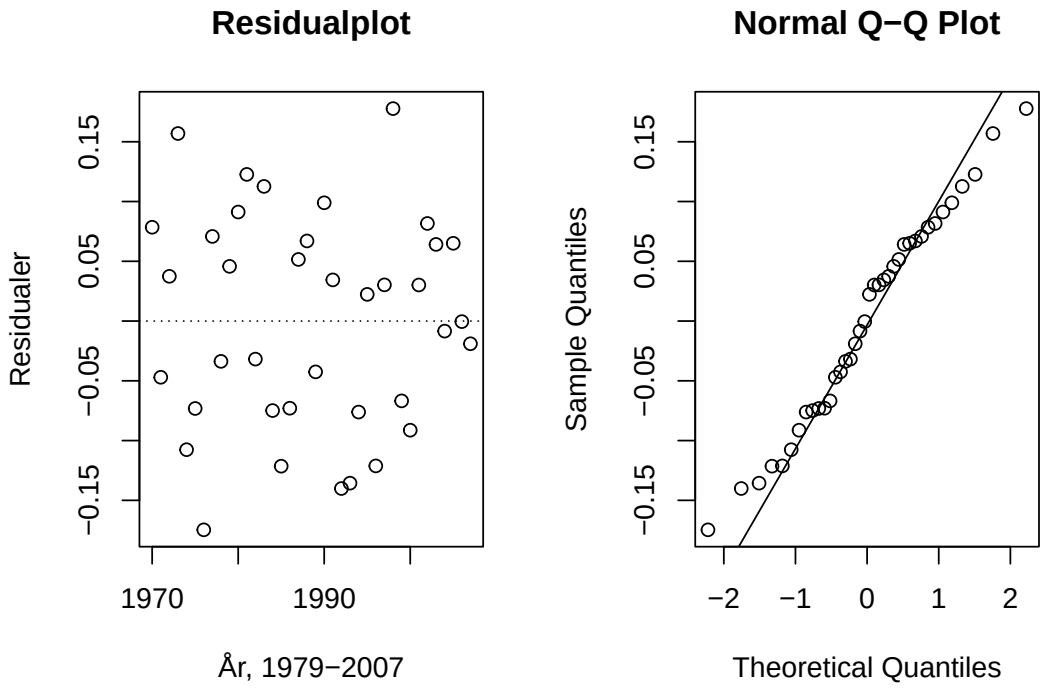
```r
# delar in i tre årsperioder
df1 <- subset(df, år >= 1880 & år <= 1929)
df2 <- subset(df, år >= 1930 & år <= 1969)
df3 <- subset(df, år >= 1970 & år <= 2007)

# 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)
```



```r
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)
```




```r
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.

```
summary1 <- summary(modell1)
summary2 <- summary(modell2)
summary3 <- summary(modell3)

# extrahera alpha
coef1 <- coef(modell1)
coef2 <- coef(modell2)
coef3 <- coef(modell3)

# extrahera beta
intercept1 <- coef1[1]
slope1 <- coef1[2]

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

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


cat("Model 1: Intercept =", intercept1, ", Lutningskoefficient =", slope1, "\n")
```

```
## Model 1: Intercept = 1.284978 , Lutningskoefficient = -0.0008419208
```

```r
cat("Model 2: Intercept =", intercept2, ", Lutningskoefficient =", slope2, "\n")
```

```
## Model 2: Intercept = -0.9199713 , Lutningskoefficient = 0.0004318011
```

```r
cat("Model 3: Intercept =", intercept3, ", Lutningskoefficient =", slope3, "\n")
```

```
## Model 3: Intercept = -34.62775 , Lutningskoefficient = 0.01752653
```

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.

```r
p_value <- summary(modell3)$coefficients[2, 4] / 2
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

Upgift 3.1

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

```r
# Modell 1: Data från 1850-2007
modell1 <- lm(temperatur ~ år, data = df[which(df$år <= 2007), ])

# Modell 2: Data från 1970-2007
modell2 <- lm(temperatur ~ år, data = df[which(df$år >= 1970 & df$år <= 2007), ])

# Generera prediktioner för åren 2008-2022
pred1 <- predict(modell1, newdata = df_test)
pred2 <- predict(modell2, newdata = df_test)

# Beräkna RMSE för åren 2008-2022
rmse1 <- sqrt(mean((df_test$temperatur - pred1)^2))
rmse2 <- sqrt(mean((df_test$temperatur - pred2)^2))
```

```r
# Filtrera testdatan för perioden 1970-2007
df_test_1970_2007 <- df_test[which(df_test$år >= 1970 & df_test$år <= 2007), ]

# Generera prediktioner för perioden 1970-2007
pred1_1970_2007 <- predict(modell1, newdata = df_test_1970_2007)
pred2_1970_2007 <- predict(modell2, newdata = df_test_1970_2007)

# Beräkna RMSE för perioden 1970-2007
rmse1_1970_2007 <- sqrt(mean((df_test_1970_2007$temperatur - pred1_1970_2007)^2))
rmse2_1970_2007 <- sqrt(mean((df_test_1970_2007$temperatur - pred2_1970_2007)^2))
```

```r
# Sammanställ alla RMSE-värden i en tabell
rmse_tabell <- data.frame(
  Modell = c("1850-2007", "1970-2007", "1850-2007", "1970-2007"),
  Testperiod = c("2008-2022", "2008-2022", "1970-2007", "1970-2007"),
  RMSE = c(rmse1, rmse2, rmse1_1970_2007, rmse2_1970_2007)
)

# Skriv ut tabellen
print(rmse_tabell)
```

```
##      Modell Testperiod      RMSE
## 1 1850-2007  2008-2022 0.5377942
## 2 1970-2007  2008-2022 0.1506799
## 3 1850-2007  1970-2007       NaN
## 4 1970-2007  1970-2007       NaN
```

Uppgift 3.2