Statistisk Analys: Laboration 3

Ottilia Andersson & David Sermoneta

2025-01-04

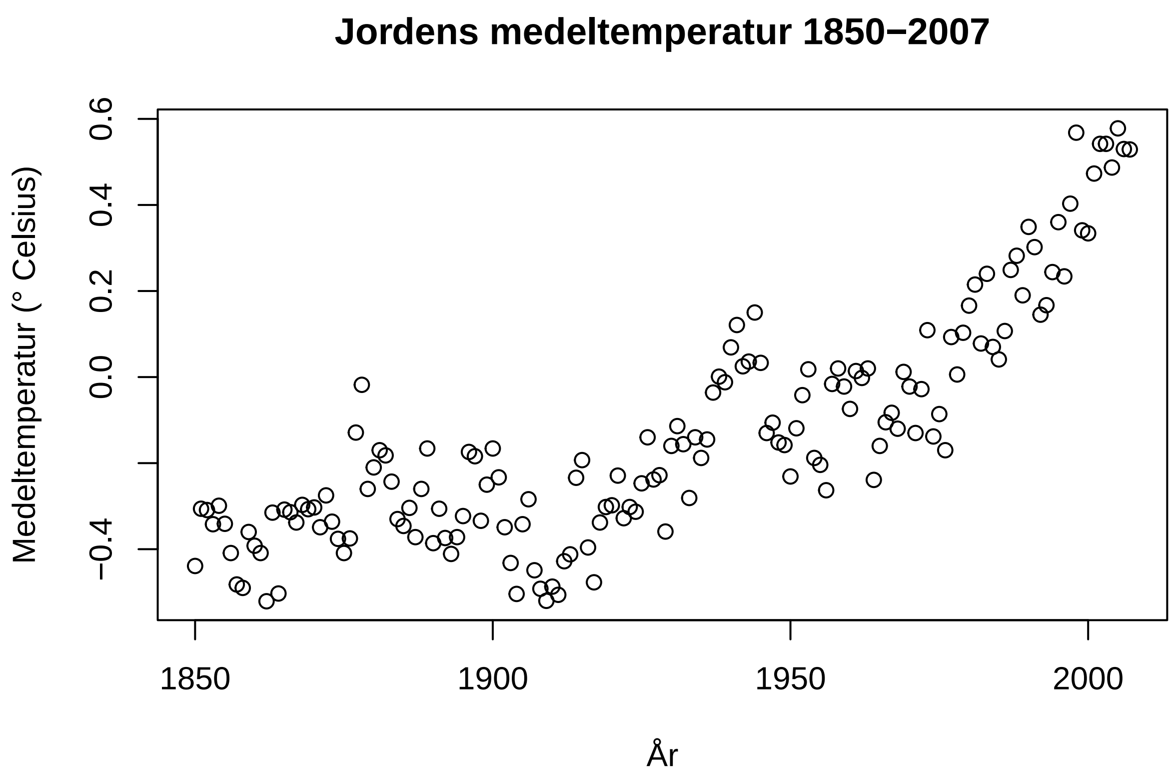
Uppgift 1: Jordens medeltemperatur 1850-2007

Vi läser in fil med information om jordens medeltemperatur åren 1850-2007.

```
df <- read.csv("temperatur.csv", header = TRUE) # Läser in jordens medeltemperatur 1850-2007
```

Vi visualiserar vår inlästa data i en scatterplot för att få en överblick över den:

```
plot(df$år,
     df$temperatur,
     main = "Jordens medeltemperatur 1850-2007",
     xlab = "År",
     ylab = "Medeltemperatur (° Celsius)")
```



Figur 1: Jordens medeltempteratur 1850-2007. År på x-axeln och temperatur på y-axeln, mätt i grader Celcius.

Det ser absolut ut att finnas ett samband mellan år och jordens medeltemperatur. Vi ser att jordens medeltemperatur med åren har stigit. Men hur ska ökningen bäst beskrivas matematiskt? Är det en linjär ökning? Om vi begränsar oss till år 1850-1975 (på ett ungefär) ser sambandet ut att vara hyfsat linjärt. Detsamma gäller tidsperioden 1975-2007, men här är ökar temperaturen snabbare, vilket vi ser i form av att lutnignen är brantare. Betraktar vi hela tidsperioden kan vi alltså - med tanke på att det ser ut som att temperaturen har ökat olika snabbt under olika tidsperioder - få problem om vi vill beskriva sambandet med en linjär modell. Man skulle kunna tänka sig att sambandet mellan tid och temperatur snarare är exponentiellt. Detta återstår att undersöka vidare!

Men om vi ändå vill försöka anpassa en linjär modell till datamaterialet, skulle vi över huvud taget kunna utföra enkel linjär regression?

För att kunna utföra enkel linjär regression behöver vi ha obseroende observationer, dvs att temperaturen vid mätning ett år inte beror på temperaturen vid mätning ett annat år. Detta kan vi anta är uppfyllt. Vi behöver bivariat data, med en förklarande variabel som inte beror på slump (det vi "kontrollerar") och en responsvariabel som beror på slump (det vi "mäter"). Detta är uppfyllt i vårt material: *år* är en förklarande variabel och *temperatur* är en responsvariabel.

Vi behöver att det råder ett linjärt samband mellan den förklarande variabeln och responsvariabeln. Som tidigare nämnt råder viss tvekan gällande linjaritet över hela tidsperioden. Utifrån okulär besiktning av data (dvs vi tittar på vår scatterplot och gör en ungefärlig bedömning), tänker vi att det bästa vore att dela upp materialet i två tidsperioder, med brytpunkt kring år 1975, om vi vill utföra enkel linjär regression. Vi kan dock testa och se hur välanpassad en linje blir över hela datamaterialet.

Därutöver behöver vi att residualerna är normalfördelade, samt att variansen i residualerna är konstant. (Vi återkommer till detta längre ner.)

Vi använder `R:s`´inbuggda funktion för linjär regression:

```
modell <- lm(temperatur ~ 1 + år, data = df) # Anpassar rät linje med temperatur som responsvariabel oc
```

Vi visualiserar vår anpassade linje i vår scatterplot igen:

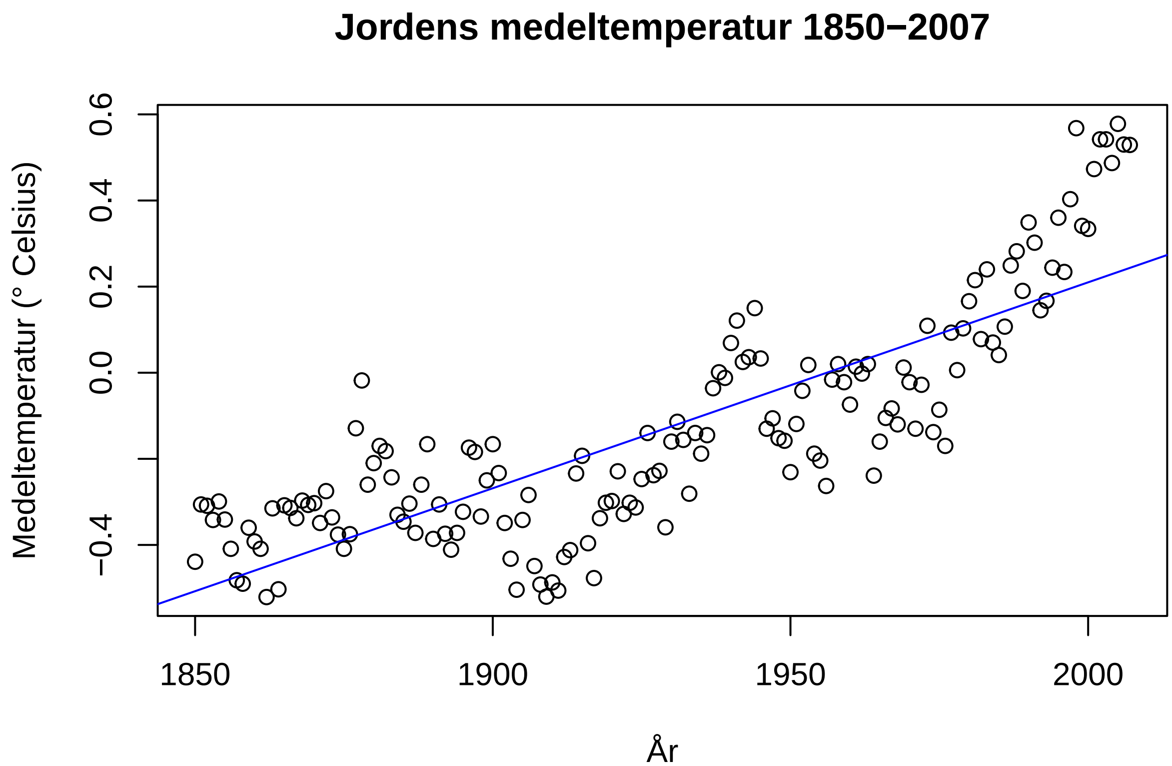
```
plot(df$år, # Plottar datamaterialet likt tidigare
     df$temperatur,
     main = "Jordens medeltemperatur 1850-2007",
     xlab = "År",
     ylab = "Medeltemperatur (° Celsius)")
abline(modell, # Lägger till anpassad linje
       col = "blue")
```

Det är svårt att med blotta ögat avgöra hur väl linjen passar. Vi kan skönja ett mönster för de punkter som ligger utanför linjen, nämligen att det finns fler punkter *ovan* linjen ute på kanterna, dvs tidiga respektive sena årtal, medan det i mitten av datamaterialet, dvs omkring år 1900-1980, finns fler punkter *under* linjen. Detta är helt i linje med våra kommentarer ovan, att det finns en exponentiell tendens i datamaterialet, sett över hela tidsperioden.

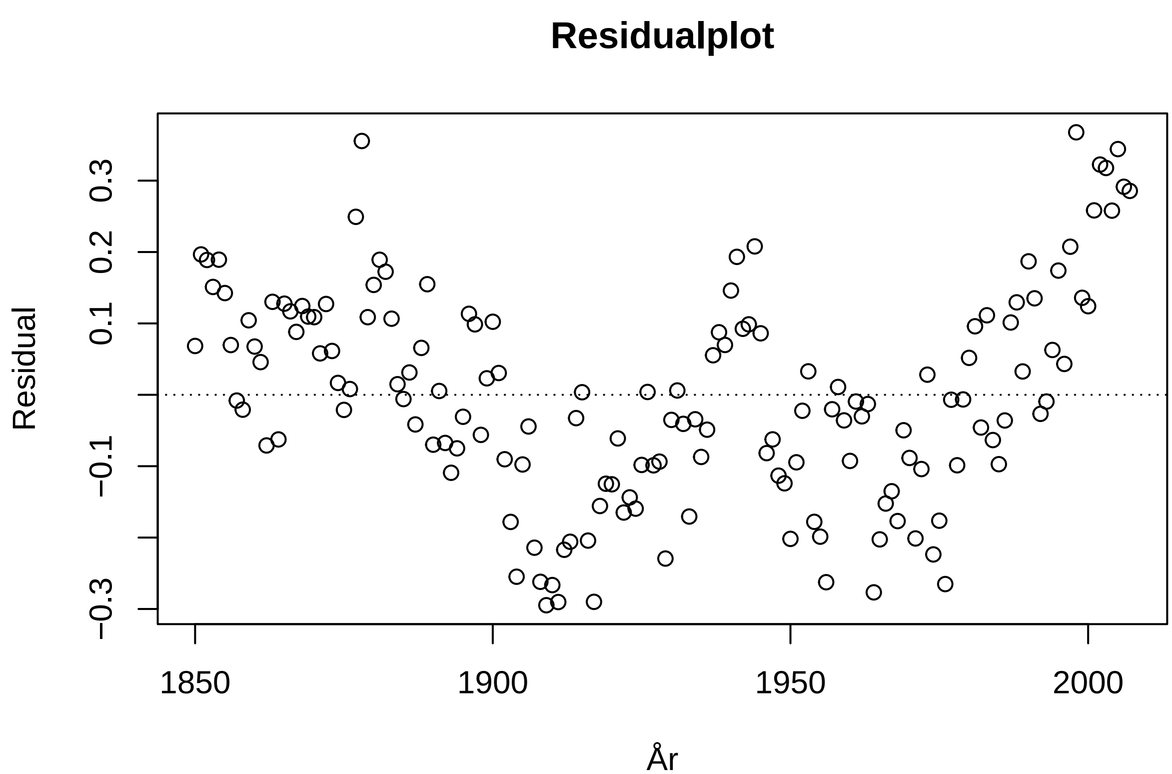
Två hjälpmedel för att undersöka hur väl linjen passar datamaterialet är residualplot och normalfördelningsplot. Residualer mäter individuella observationers avvikelse från den anpassade linjen och residualplotten hjälper oss således att avgöra om en linjär modell är lämplig för våra data. Vi undersöker:

```
residual <- modell$residuals # Extraherar residualer
plot(df$år, residual, # Skapar residualplot
     main = "Residualplot",
     xlab = "År",
     ylab = "Residual")
abline(a = 0, b = 0, lty = "dotted") # Lägger till horisontell linje
```

I figuren ovan kan vi skönja en trend bland residualerna. De är inte jämnt spridda, utan de första decennierna ser vi en nedåtgående trend, och de sista decennierna ser vi i stället en uppåtgående trend. Vi kan också se olika stora avvikelser från den horisontella mittlinjen för olika segment längs x-axlen. Detta är indikationer



Figur 2: Jordens medeltempteratur 1850-2007. År på x-axeln och temperatur på y-axeln, mätt i grader Celcius. Anpassad regressionslinje i blått.

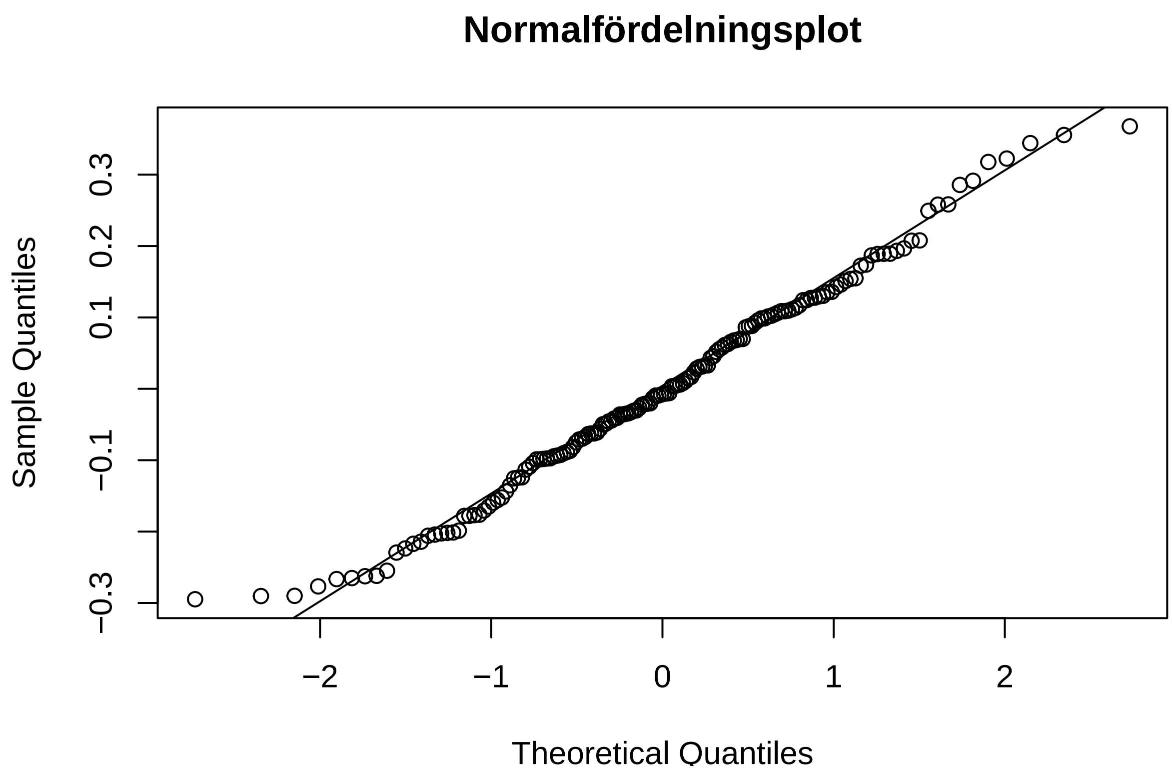


Figur 3: Residualplot över jordens medeltemperatur, tidigare plottad.

på att variansen inte är konstant, och en linjär modell kanske inte är den bästa för vårt datamaterial. Dessa kriterier för enkel linjär regression (som vi diskuterade ovan) verkar med andra ord inte helt vara uppfyllda.

Vi går vidare och betraktar en normalfördelningsplot:

```
qqnorm(residual,
        main = "Normalfördelningsplot")
qqline(residual)
```



Figur 4: En normalfördelningsplot över residualerna från tidigare plottad linjär regression över jordens medeltemperatur.

Normalfördelningsplotten säger oss hur spridningen bland residualerna är. Som tidigare nämnt vill vi att residualerna ska vara normalfördelade. I figuren ovan ser vi att de allra flesta punkterna ligger längs en rät linje, och utifrån det kan vi sluta oss till att residualerna är normalfördelade. För att knyta an till vår tidigare diskussion om förutsättningar för enkel linjär regressoin, kan vi alltså konstatera att kriteriet om normalfördelade residualer är uppfyllt.

Uppgift 2

Nu gör vi samma sak fast vi delar in perioden i tre delperioder!

```
# Dela upp datan i 3 perioder
df_1 <- subset(df, år > 1879 & år < 1930)
df_2 <- subset(df, år > 1929 & år < 1970)
df_3 <- subset(df, år > 1969)

# Skapa tre modeller för varje årsperiod
modell_1 <- lm(temperatur ~ 1 + år, data = df_1)
modell_2 <- lm(temperatur ~ 1 + år, data = df_2)
```

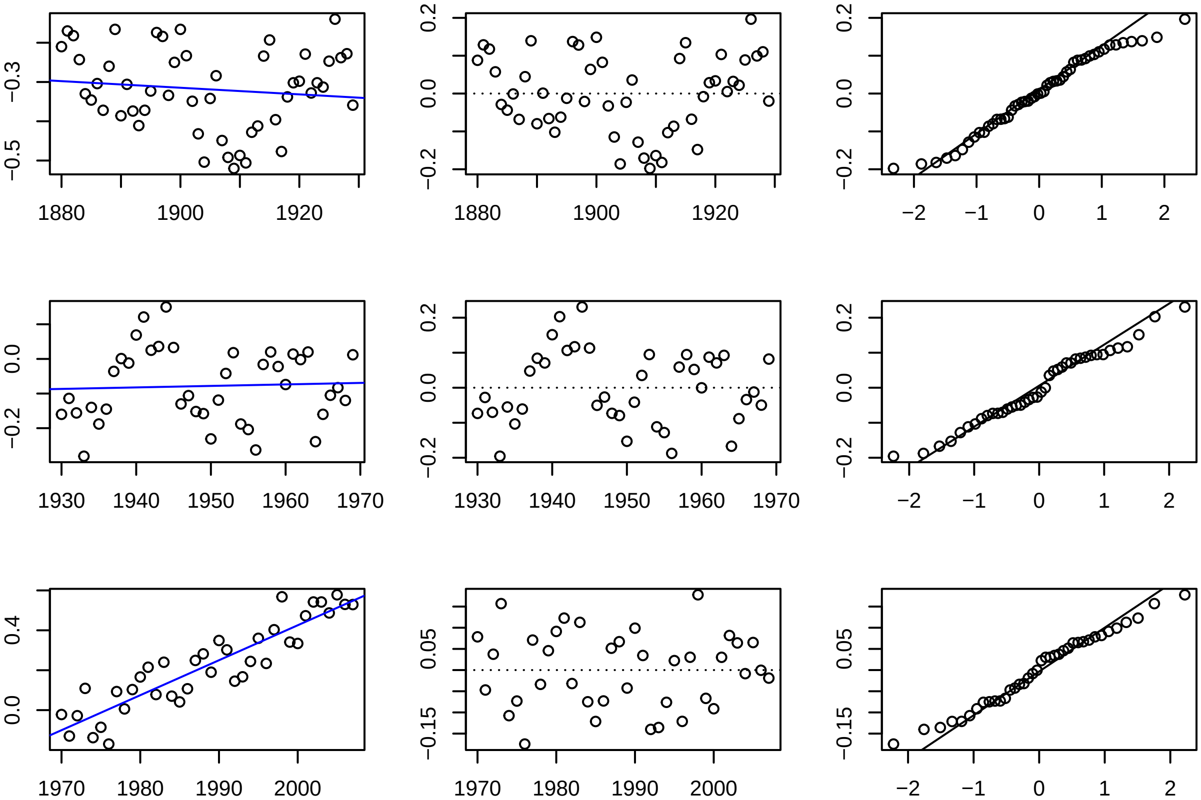
```r
modell_3 <- lm(temperatur ~ 1 + år, data = df_3)
```

Vi vill undersöka om dessa tre perioder uppfyller förutsättningarna för enkel linjär regression. I figur 5 ser vi att första och andra perioden inte har så bra förutsättningar för linjär regression. Residualerna är inte alls särskilt jämnt spridna. För den tredje är detta inte fallet, där ser vi en jämn spridning av residualerna. Normalfördelningsplottarna ger vidare att alla tre perioder verkar ha normalfördelade residualer. Resten av förutsättningarna ärver vi från den större perioden.

```r
# Set smaller margins
par(mfrow = c(3, 3), mar = c(3, 3, 2, 1))  # c(bottom, left, top, right)

# Plotta alla punkter i årsperioden över årtalen
plot(df_1$år, df_1$temperatur, xlab = "År", ylab = "Temperatur")
abline(modell_1, col = "blue")

# Plotta Residualer
plot(df_1$år, modell_1$residuals, xlab = "År", ylab = "Residualer")
abline(a = 0, b = 0, lty = "dotted")

qqnorm(modell_1$residuals, main = "")
qqline(modell_1$residuals)



plot(df_2$år, df_2$temperatur, xlab = "År", ylab = "Temperatur")
abline(modell_2, col = "blue")

plot(df_2$år, modell_2$residuals, xlab = "År", ylab = "Residualer")
abline(a = 0, b = 0, lty = "dotted")

qqnorm(modell_2$residuals, main = "")
qqline(modell_2$residuals)


plot(df_3$år, df_3$temperatur, xlab = "År", ylab = "Temperatur")
abline(modell_3, col = "blue")

plot(df_3$år, modell_3$residuals, xlab = "År", ylab = "Residualer")
abline(a = 0, b = 0, lty = "dotted")

qqnorm(modell_3$residuals, main = "")
qqline(modell_3$residuals)
```

Nu utför vi en enkel linjär regression på alla tre perioder, samt sparar skattningarna vi får av α och β för alla tre perioder, samt p-värdena för alla dessa för skojs skull. Vi visar upp detta

```r
sammanfattning_1 <- summary(modell_1)
sammanfattning_2 <- summary(modell_2)
sammanfattning_3 <- summary(modell_3)

alpha1 <- c(coef(sammanfattning_1)[1], coef(sammanfattning_1)[1, "Pr(>|t|)"])
beta1 <- c(coef(sammanfattning_1)[2], coef(sammanfattning_1)[2, "Pr(>|t|)"])

alpha2 <- c(coef(sammanfattning_2)[1], coef(sammanfattning_2)[1, "Pr(>|t|)"])
beta2 <- c(coef(sammanfattning_2)[2], coef(sammanfattning_2)[2, "Pr(>|t|)"])
```



Figur 5: Diverse plottar av regressionslinje (vänster), residualplott (mitten), och normalfördelningsplott (höger). Detta för alla periodindelningar 1880 - 1929, 1930 - 1969, och 1970 - 2007.

```r
alpha3 <- c(coef(sammanfattning_3)[1], coef(sammanfattning_3)[1, "Pr(>|t|)"])
beta3 <- c(coef(sammanfattning_3)[2], coef(sammanfattning_3)[2, "Pr(>|t|)"])

table <- data.frame(perioder = c("1880-1929", "1930-1969", "1970-2007"),
                    alpha = c(alpha1[1], alpha2[1], alpha3[1]),
                    beta = c(beta1[1], beta2[1], beta3[1]),
                    pvärdeb = c(beta1[2], beta2[2], beta3[2]))

colnames(table) <- c("Årsperiod", "alpha",
                   "beta", "p-värde för hypotesen beta = 0")

knitr::kable(table, caption = "Tabell av värden på skattningar av alpha, beta, och p-värden för betas sk
```

Tabell 1: Tabell av värden på skattningar av alpha, beta, och p-värden för betas skattningar.

Årsperiod	alpha	beta	p-värde för hypotesen beta = 0
1880-1929	1.2849781	-0.0008419	0.4104753
1930-1969	-0.9199713	0.0004318	0.7693833
1970-2007	-34.6277510	0.0175265	0.0000000

Som vi ser i tabell 1, så är skattnnigarna av lutningskoefficienterna (β) rätt så usla (högt p-värde) för första två perioderna, vilket stärker våran uppfattning om att dessa två årsperioderna inte var lämpade till linjär regression någorlunda. Medans skattningarna för α och β i tredje perioden var rätt så toppen, vilket tar oss vidare till nästa del.

Uppgift 2.2

Här testar vi direktörens hypotes om att det inte finns en trend mot ett varmare klimat, dvs, $H_0 : \beta \leq 0$. Då blir vår alternativa hypotes $H_1 : \beta > 0$. Vi väljer att testa β eftersom våran förklarande variabel inte är stokastisk, dvs, vi vill bara se om år, som ej är stokastisk, har en positiv effekt på medeltemperatur. Vi utför alltså ett ensidigt t-test, eftersom vi försöker skatta en parameter β av data, som tack vare centrala gränsvärdessatsen, tillhör en normalfördelning med okänd varians.

Anledningen till att vi utför ett ensidigt hypotestest är ju att det ingår i direktörens påstående att det skulle till och med kunna finnas en negativ trend mellan år och temperatur, dvs, att världen blir kallare, så det känns rimligt att även testa detta.

Går vi tillbaka till tabell 1, ser vi att β för årsperioden $1970 - 2007$ är otroligt liten, men ändå nollskillt! Vidare har vi ett extremt litet p-värde från vårat t-test (i princip noll), som vid halvering (eftersom `summary()` endast ger ett tvåsidigt test, men detta är inte ett problem eftersom responsvariabeln är normalfördelad, dvs symmetrisk), blir ännu mindre. Detta är (sjukt nog) mindre än 0.05, som är den signifikansnivå som angavs i instruktionerna, och därför kan vi med god säkerhet förkasta nollhypotesen, och dra slutsatsen om att det finns en *tydlig* trend mot ett varmare klimat mellan åren 1970 och 2007. Med andra ord hade direktören fel.

Upgift 3.1

För att ens kunna börja besvara på frågorna ovan behöver vi först definiera funktionen som ger oss medelkvadratfelet:

```r
#Funktion som ger roten r medelkvadratfelet (RMSE) mellan predikterade och faktiska värden, roten ur fö
rmse <- function(y_faktiska, y_pred){
  diff <- y_faktiska - y_pred
```

```r
  sqr_diff <- diff^2
  mse <- mean(sqr_diff)
  return(sqrt(mse))
}
```

Nu kan vi börja göra lite prediktioner! Vi börjar med att läsa in de faktiska värdena för perioden 2008 – 2022

```r
# Låt period 1 ("p1") beteckna årsindelningen 2008-2022,
# och period 2 ("p2") beteckna 1970 - 2007.
# Vidare betecknar "hela", årsperioden 1850 - 2007

test_df <- read.csv("temperatur_test.csv", header = TRUE)

x_p1 <- test_df$år # sekvens med åren 2007-2022

df_p1 <- data.frame(år = x_p1)  # Vi gör en speciell vektor med åren som vi kan använda oss av
```

```r
# Prektionsmodell för hela årsperioden 1850 - 2007 (hela), av 2008-2022 (p1)
prediktion_hela_p1 <- predict(modell, newdata = df_p1, interval = 'predict')

# Plocka ut prediktioner samt gränser på prediktionsintervallet
y_pred_hela_p1 <- prediktion_hela_p1[, 1]

#Prediktionsmodell för 1970 - 2007 (p2), av 2008 - 2022 (p1)
prediktion_p2_p1 <- predict(modell_3, newdata = df_p1,
                             interval = 'predict')

y_pred_p2_p1 <- prediktion_p2_p1[, 1]
pinterval_undre_p1 <- prediktion_p2_p1[, 2]
pinterval_övre_p1 <- prediktion_p2_p1[, 3]
```

```r
# Här gör vi prediktionerna som ovan, fast av perioden 1970 - 2007
x_p2 <- df_3$år

df_p2 <- data.frame(år = x_p2)

#Prediktionsmodell för alla år (hela) av 1970 - 2007 (p2)
prediktion_hela_p2 <- predict(modell, newdata = df_p2, interval = 'predict')

y_pred_hela_p2 <- prediktion_hela_p2[, 1]

#Prediktionsmodell för åren 1970-2007 (p2) av 1970-2007 (p2)
prediktion_p2_p2 <- predict(modell_3, newdata = df_p2, interval = 'predict')

y_pred_p2_p2 <- prediktion_p2_p2[, 1]
```

Vi noterar några saker från tabell 2. Först, att prediktionerna för medeltemperaturen åren 2008-2022 görs bäst av den justerade modellen (dvs, den från 1970-2007). Detta förstärker vår idé om att hela perioden är inget att ha. En till sak är att när en modell predikterar samma värden som den utgörs av, får vi ett väldigt litet, (men ändå nollskillt!) RMSE. Men ännu en grej som får oss att tycka att vår allra första modell är kass som en linjär regressionsmodell, är att när den försöker prediktera en delmängd av årperioden som den utgörs av, så gör den ett *sämre* jobb (dvs får ett högre RMSE), än vad den justerade modellen lyckas med på helt nya värden!

Med detta sagt, är det fortfarande rätt stora fel vi finner, (som även figur 6 delvist säger oss). Även om talen är små, är det värt att påminna läsaren att det vi undersöker är rätt så små temperaturskillnader, så ett snittligt fel på ca 0.15 är rätt så stort när man tänker på att datan vi undersöker har skillnader i temperatur så små som 0.05 (t.ex. mellan åren 1985 till 1986).

```r
y_faktiska_p1 <- test_df$temperatur

rmse_hela_p1 <- rmse(y_faktiska_p1, y_pred_hela_p1)
rmse_p2_p1 <- rmse(y_faktiska_p1, y_pred_p2_p1)

y_faktiska_p2 <- df_3$temperatur
rmse_hela_p2 <- rmse(y_faktiska_p2, y_pred_hela_p2)
rmse_p2_p2 <- rmse(y_faktiska_p2, y_pred_p2_p2)

table2 <- data.frame(
  anpassning = c("1850 - 2007", "1970 - 2007", "1850 - 2007", "1970 - 2007"),
  pred = c("2008 - 2022", "2008 - 2022", "1970 - 2007", "1970 - 2007"),
  rmse_värden = c(rmse_hela_p1, rmse_p2_p1, rmse_hela_p2, rmse_p2_p2)
)

colnames(table2) <- c("Modell anpassad för", "Prediktioner för", "RMSE")
knitr::kable(table2, caption = "Tabell med RMSE för olika modeller och deras prediktioner")
```

Tabell 2: Tabell med RMSE för olika modeller och deras prediktioner

Modell anpassad för	Prediktioner för	RMSE
1850 - 2007	2008 - 2022	0.5377942
1970 - 2007	2008 - 2022	0.1506799
1850 - 2007	1970 - 2007	0.1788645
1970 - 2007	1970 - 2007	0.0878726

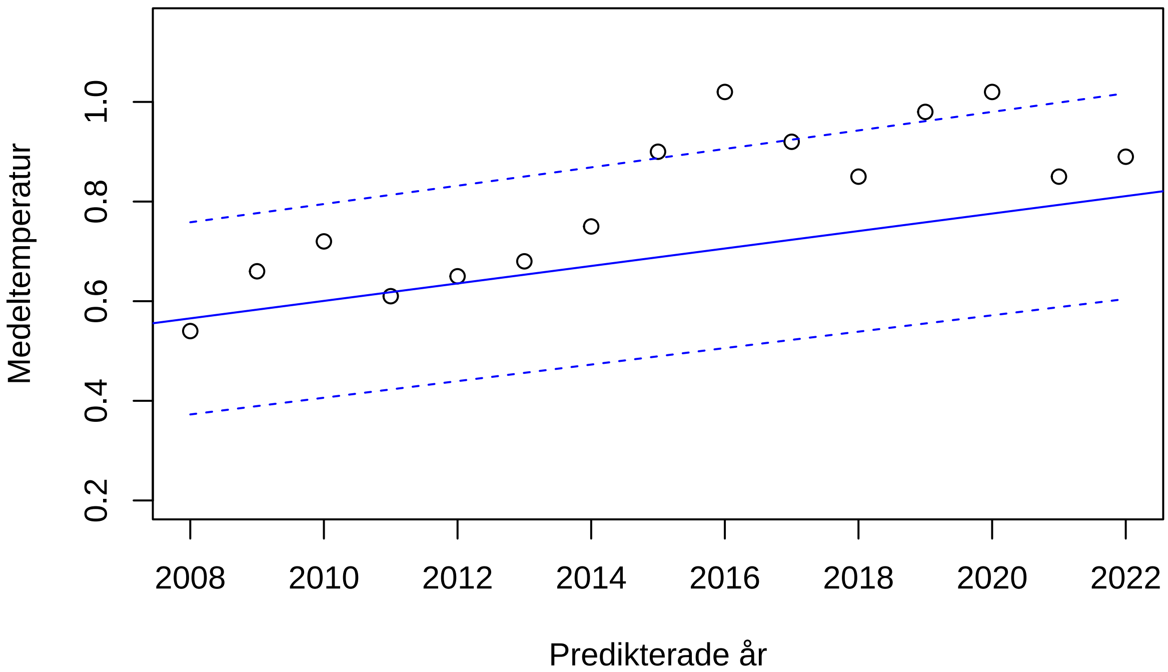
Uppgift 3.2

```r
#plottar predikterade värdena tillsammans med prediktionsintervall

plot(x_p1, y_faktiska_p1, ylim = c(0.2, 1.15),
     xlab = "Predikterade år",
     ylab = "Medeltemperatur")

lines(x_p1, pinterval_undre_p1, col = "blue",
      lty = "dashed")
lines(x_p1, pinterval_övre_p1, col = "blue",
      lty = "dashed")
abline(modell_3, col = "blue")
```

Förmodligen inte. Vi kollar endast på en förklarande variabel som är högst otrolig att kunna omfamna alla faktorer som går in i något så komplext som temperatur. Vidare ser vi i figur 6 att många av faktiska punkterna ligger utanför prediktionsintervallet. Med tanke på att ökningen i temperatur verkar ha ökat mer än linjärt på senaste tiden, så är heller inte linjaritet ett givet längre in i framtiden.

Nu vet vi inte hur långt ifrån kursinnehållet vi går, men man hade t.ex. kunnat lägga till fler förklarande variabler som en del av modellen. T.ex. genom att titta på det snittliga koldioxid-halten i atmosfären per år.

Något som är problematiskt är även att ett stort antagande görs när vi försöker prediktera framtida tempratur med dessa modeller, och det är att vi antar att allt som sker i världen hålls konstant. Det finns inte alltid bra skäl att tro att så är fallet. Som vi har sett med t.ex. nya teknologier genom tiderna så har



Figur 6: Prediktion av medeltemperaturen för åren 2008 - 2022 (blå linjen), med 95% konfidensintervall (streckade linjerna, samt faktiska värden.

dessa lett till en stor ökning av koldioxidutsläpp efter industriella revolutionen, och detta kan ju även hända igen (om det inte redan har hänt). Dessa kan ha en effekt på globala temperaturen som inte fångas upp av en sån enkel modell som våran.