lab3

Max Brehmer & Esma Özdolap

2022-01-20

Samanfattning

I denna labb kommer enkel regression och korrelationsanlys användas för att studera egenskaer hos ett material som består av bivariata data.

Uppgift 1

För uppgifterna i denna labb har filen temperatur.csv laddats ner som innehåller data över avvikelsen av jordens medeltemperatur mellan åren 1850-2007. Bland annat ska scartterplot och normalfördelningsplott skapas för att få en uppfattning om data ser linjärt ut.

```
df <- read_csv("data/temperatur.csv")
```

```
ggplot(data = df, aes(x = year, y = globe)) +
  geom_point() +
  geom_line()
```

```
ggplot(data = df, aes(x = year, y = globe)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```

```
lm_Model = lm(globe ~ year, data=df)
  Model.res = resid(lm_Model)

  plot(df$year, Model.res, ylab="Residual", xlab="År") +
  abline(0, 0)
```

```
## integer(0)
```

```
qqnorm(lm_Model$residuals)
qqline(lm_Model$residuals)
```

Enligt figur 1 ser sambandet mellan tid och global medeltemperatur ej linjär ut.

Om vi låter Y_i ha en fördelning som beror på en förklarande variabel X_i där x_i ej är en slumpvariabel, talar vi om regression. Vi kollar på en klass av modeller där observationerna är oberoende men inte lika fördelade. Modellen för linjär regressionsanalys är $Y_i = \alpha + \beta x_i + \varepsilon_i$ Där $\varepsilon_1, \varepsilon_2, ...\varepsilon_n$ är oberoende men lika fördelade med väntevärde 0 och varians är σ^2, då ett krav för linjär regresion är att variansen ska vara konstans.Oftast antar man att ε_i approximativt är normalfördelat.

I figur 2 plottas en scatterplot fast med en linjär regressionslinje till skillnad från figur 1. Vi kan i figur 2 observera att medeltemperaturen har stigit från ca -0.5 °C till 0.15 °C

I Figur 3 som är en residualplot kan vi konstatera att bandet av residualer kring 0 varierar medans spridningen

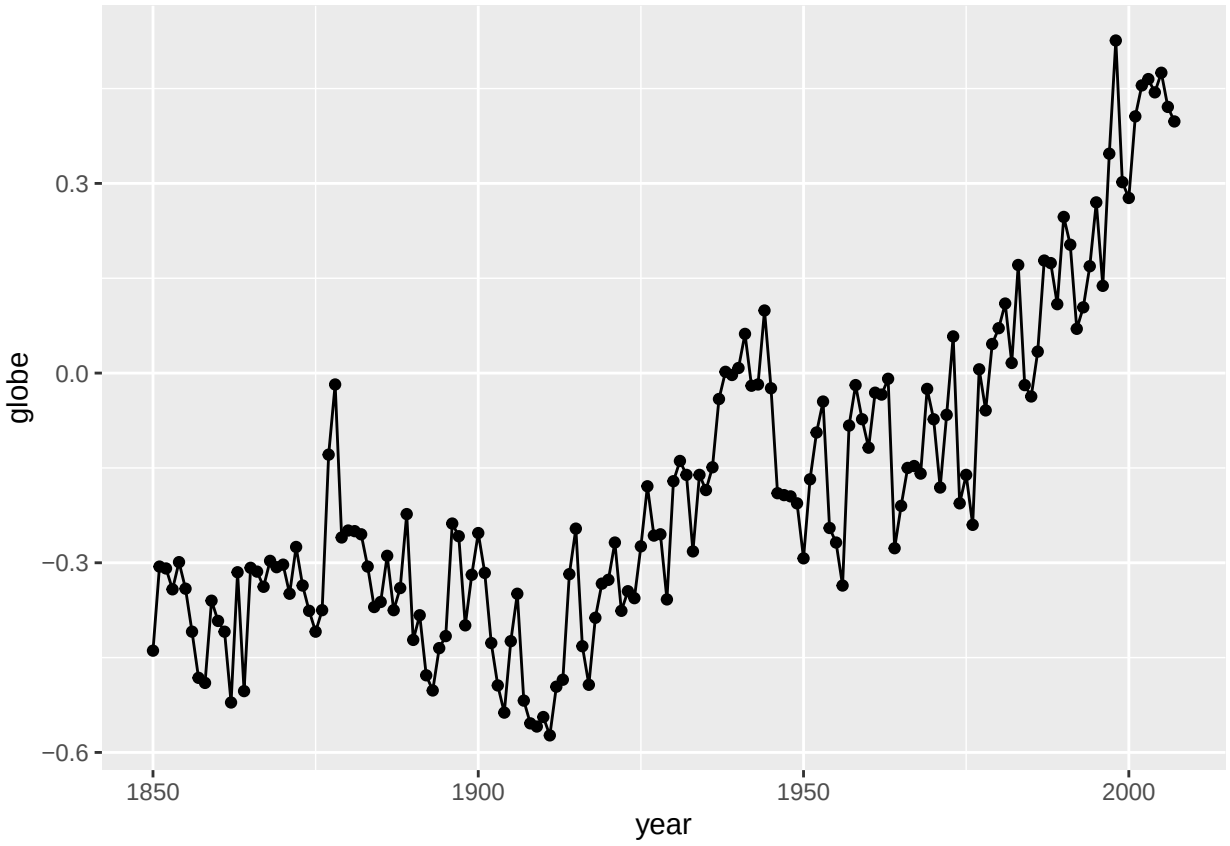

Figure 1: Temperatur över tid

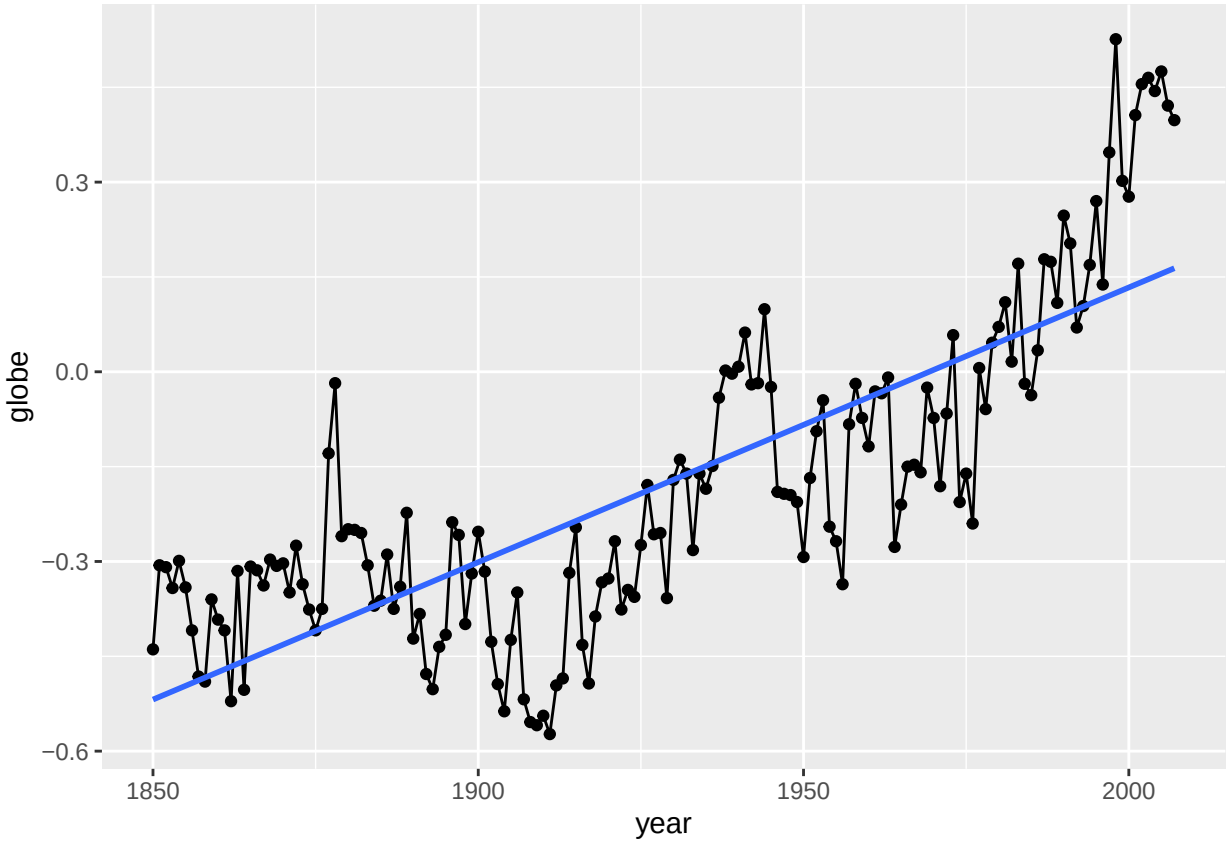

Figure 2: Linjär regression för temperatur över tid

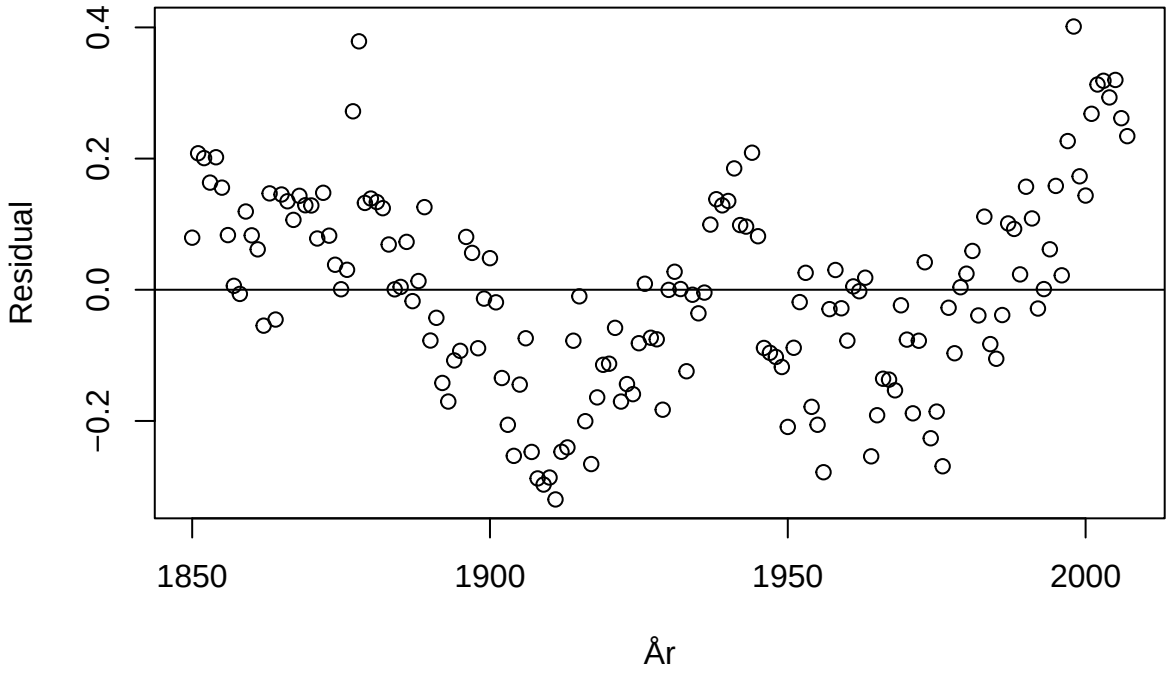

Figure 3: Residualplot över medeltemperatur och år

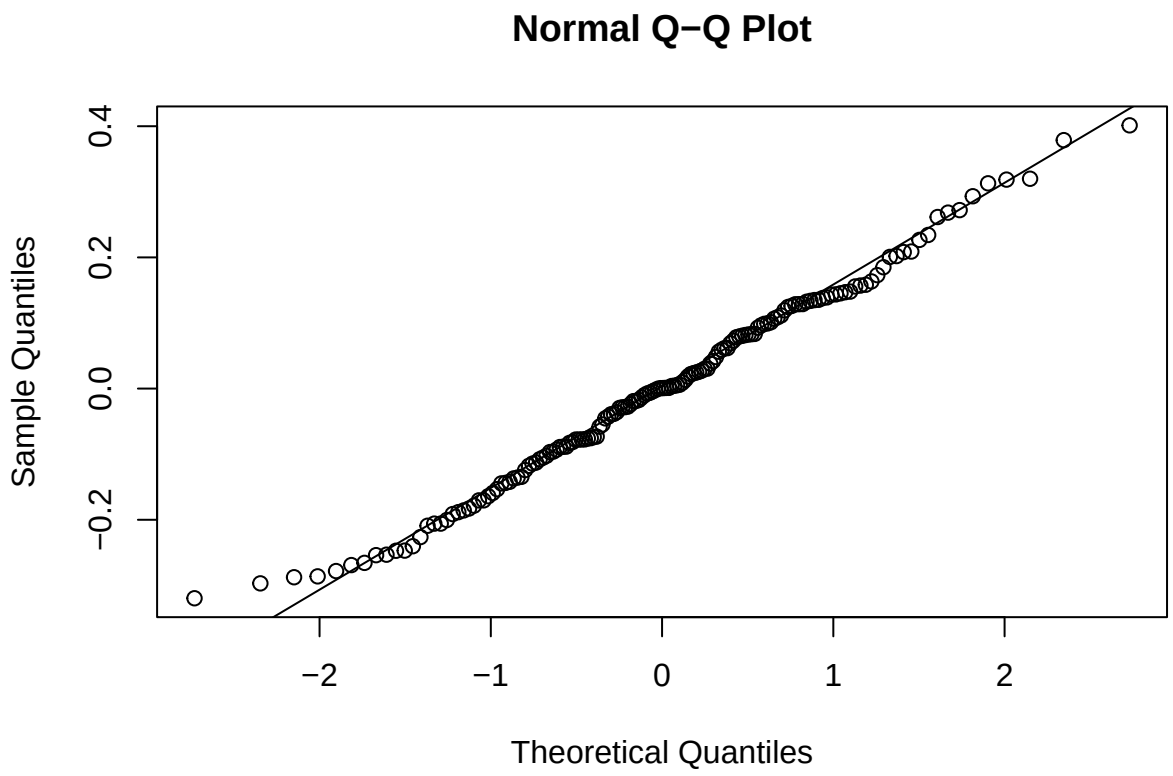

Figure 4: Normalfördelningsplot över residualen för medeltemperatur och år

inom bandet har ingen stor varians. Från figur 4 vilket är normalfördelningsplotten kan vi dra slutsatsen att residuale för global medeltemperatur och år är normalfördelad.

Slutsatsen blir att i figur 3 kan se att variansen är konstant och i figur 4 att residualerna är normalfördelade, därav uppfylls dessa två krav för linjär regression. Men eftersom relationsvariabeln beror på den förklarande variabeln uppfylls inte kravet för oberoende. Därav uppfylls inte kravet för linjär regression.

Uppgift 2

För att få modellen att passa bättre kan man dela in hela mätperioden i delperioder. I denna uppgift ska vi dela in datamaterial i tre separata tidsperioder. 1880-1929, 1930-1969 och 1979-2007.

```
df_1 <- subset(df, year >= 1880 & year <= 1929)
df_2 <- subset(df, year >= 1930 & year <= 1969)
df_3 <- subset(df, year >= 1979 & year <= 2007)
```

```
ggplot(data = df_1, aes(x = year, y = globe)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```

```
lm_Model_1 = lm(globe ~ year, data=df_1)
  Model.res = resid(lm_Model_1)

  plot(df_1$year, Model.res, ylab="Residual", xlab="År") +
  abline(0, 0)
```

```
## integer(0)
```

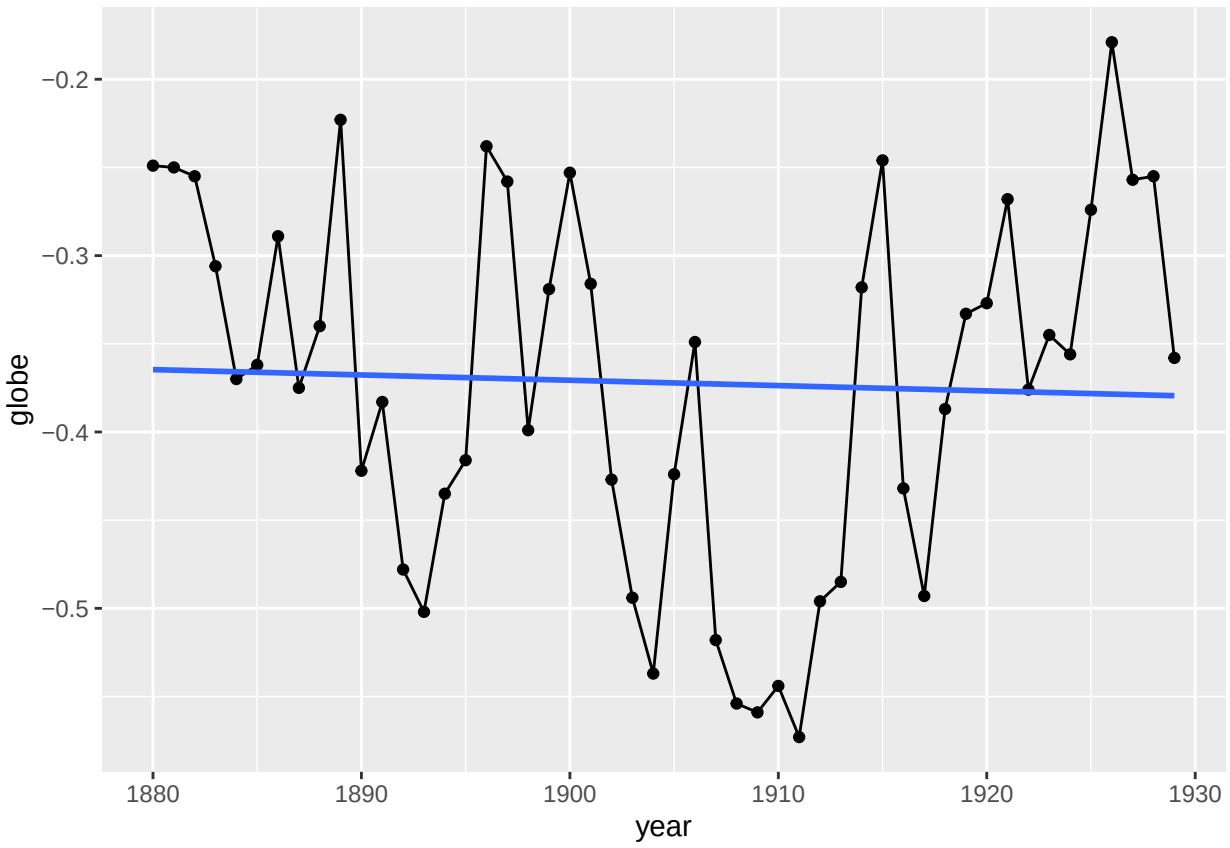

Figure 5: Linjär regression för temperatur över tid under perioden 1880-1929

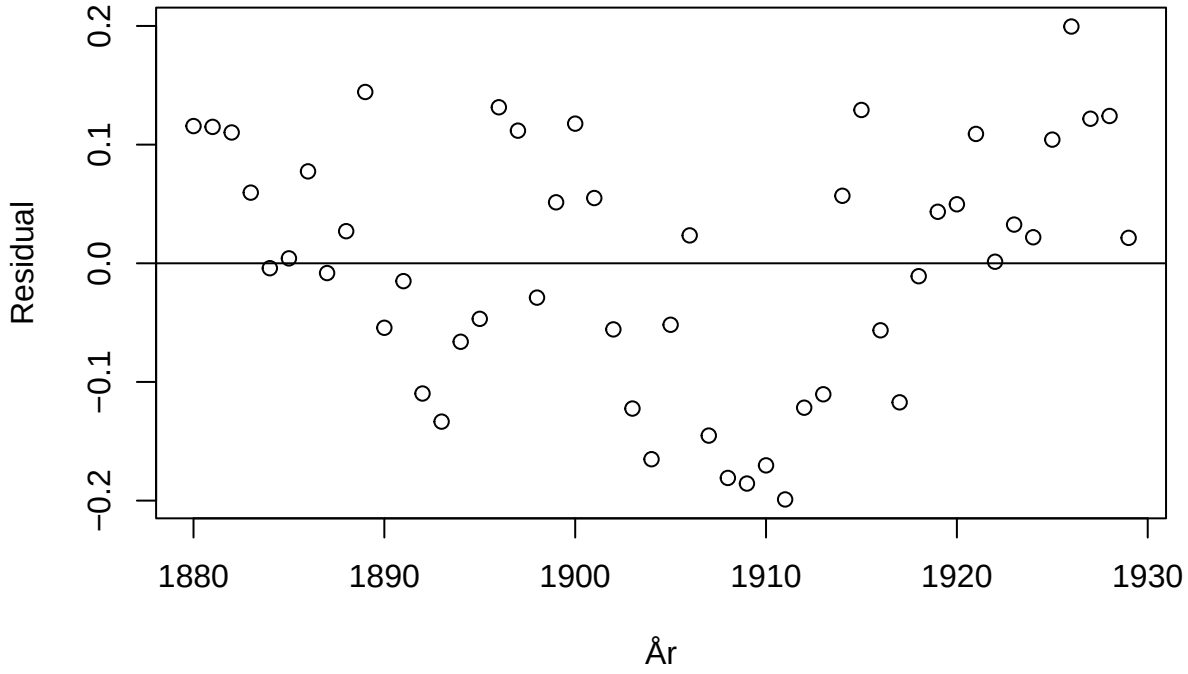

Figure 6: Residualplot över medeltemperatur och år under perioden 1880-1929

```
qqnorm(lm_Model_1$residuals)
qqline(lm_Model_1$residuals)
```

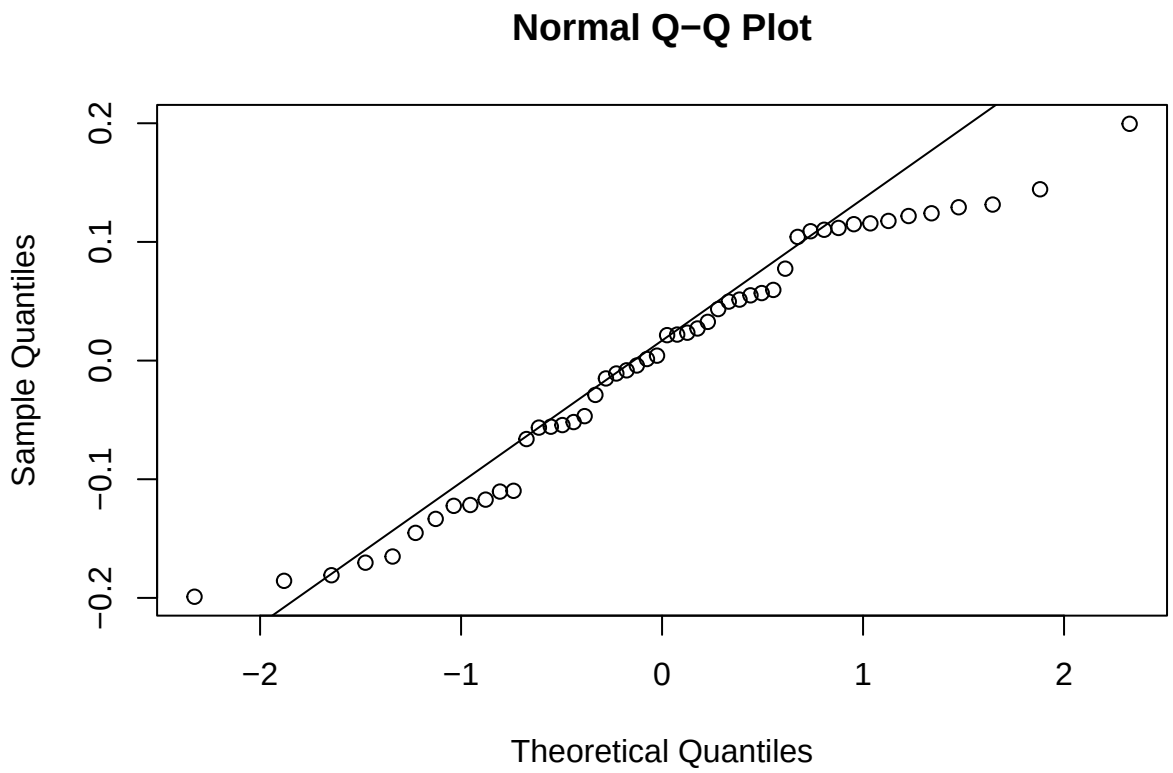

Figure 7: Normalfördelningsplot över residualen för medeltemperatur och år under perioden 1880-1929

```
ggplot(data = df_2, aes(x = year, y = globe)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```

```
lm_Model_2 = lm(globe ~ year, data=df_2)
  Model.res = resid(lm_Model_2)

  plot(df_2$year, Model.res, ylab="Residual", xlab="År") +
  abline(0, 0)
```

```
## integer(0)
```

```
qqnorm(lm_Model_2$residuals)
qqline(lm_Model_2$residuals)
```

```
ggplot(data = df_3, aes(x = year, y = globe)) +
  geom_point() +
```

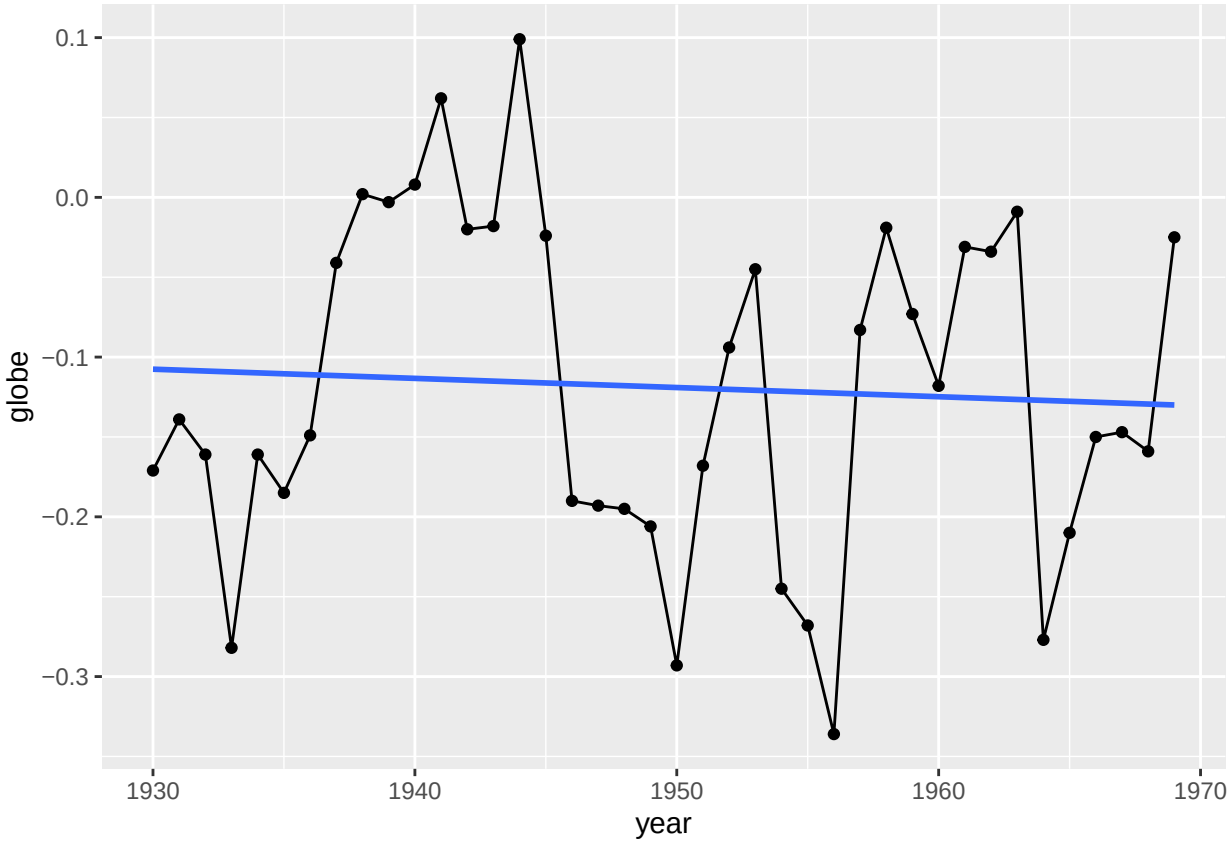

Figure 8: Linjär regression för temperatur över tid under perioden 1930-1969

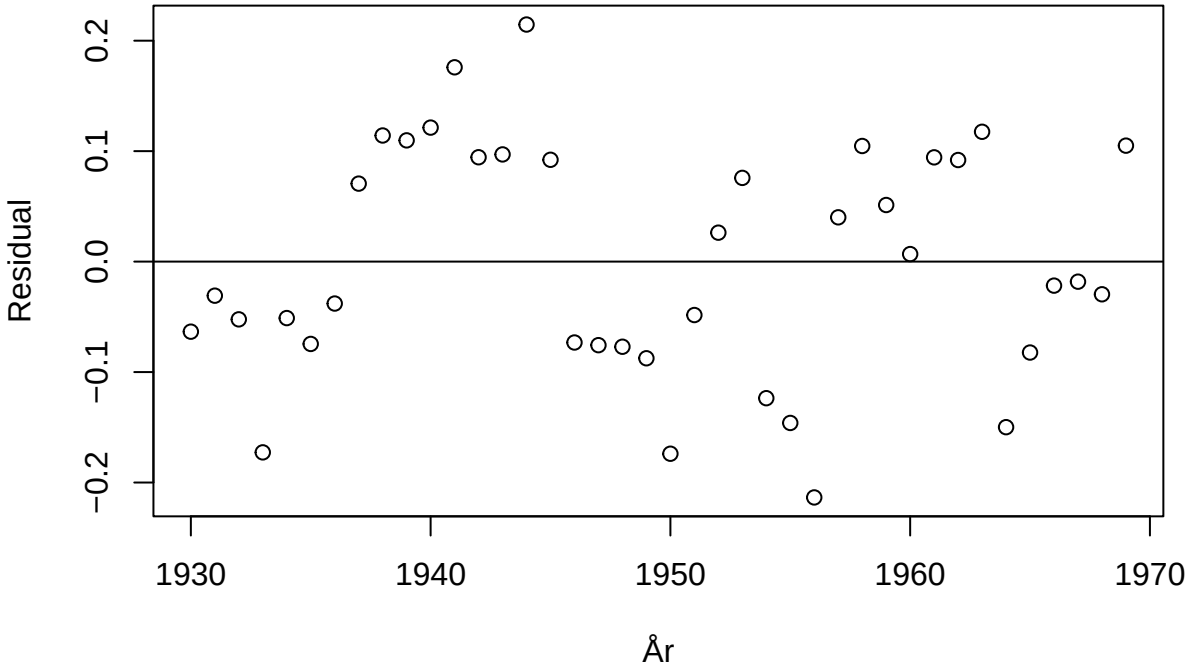

Figure 9: Residualplot över medeltemperatur och år under perioden 1930-1969

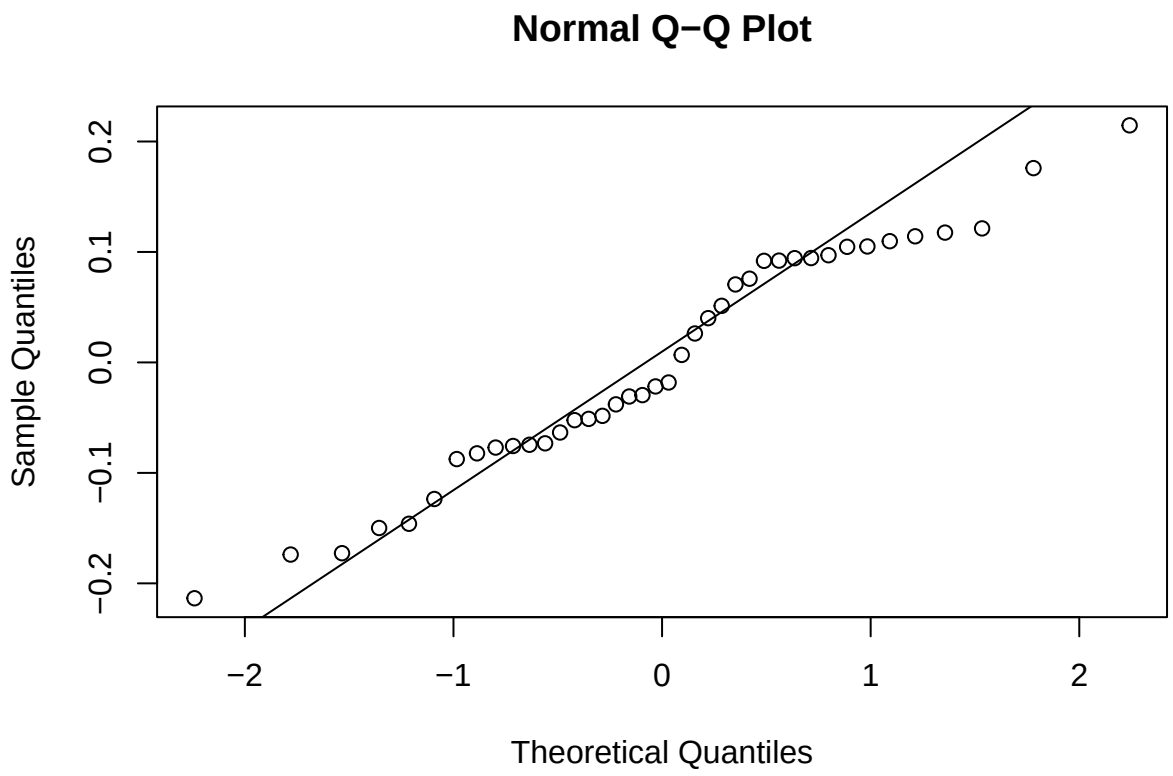

Figure 10: Normalfördelningsplot över residualen för medeltemperatur och år under perioden 1930-1969

```
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```

```
lm_Model_3 = lm(globe ~ year, data=df_3)
  Model.res = resid(lm_Model_3)

  plot(df_3$year, Model.res, ylab="Residual", xlab="År") +
  abline(0, 0)
```

```
## integer(0)
```

```
qqnorm(lm_Model_3$residuals)
qqline(lm_Model_3$residuals)
```

```
summary(lm_Model_1)
```

```
##
## Call:
## lm(formula = globe ~ year, data = df_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19900 -0.06372  0.01279  0.09753  0.19953
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2026480  1.9580586   0.103    0.918
```

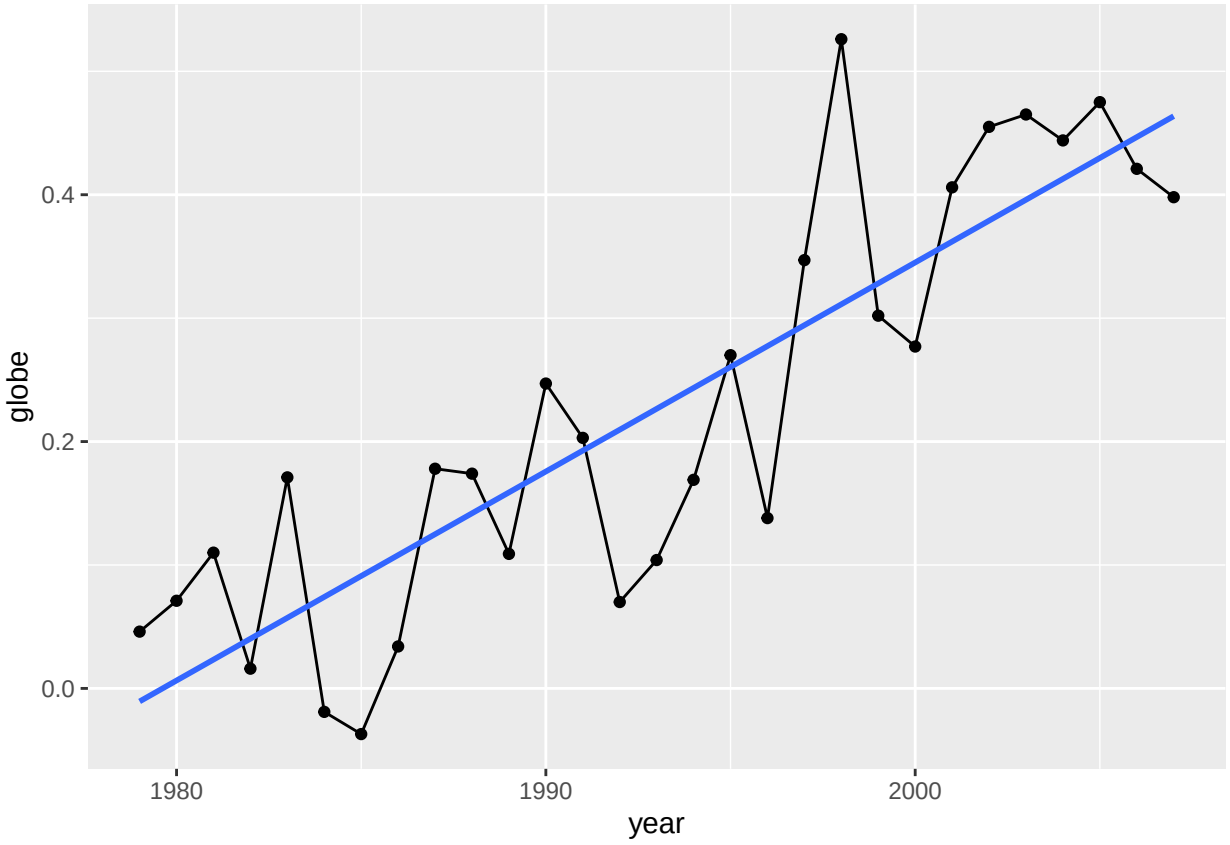

Figure 11: Linjär regression för temperatur över tid under perioden 1979-2007

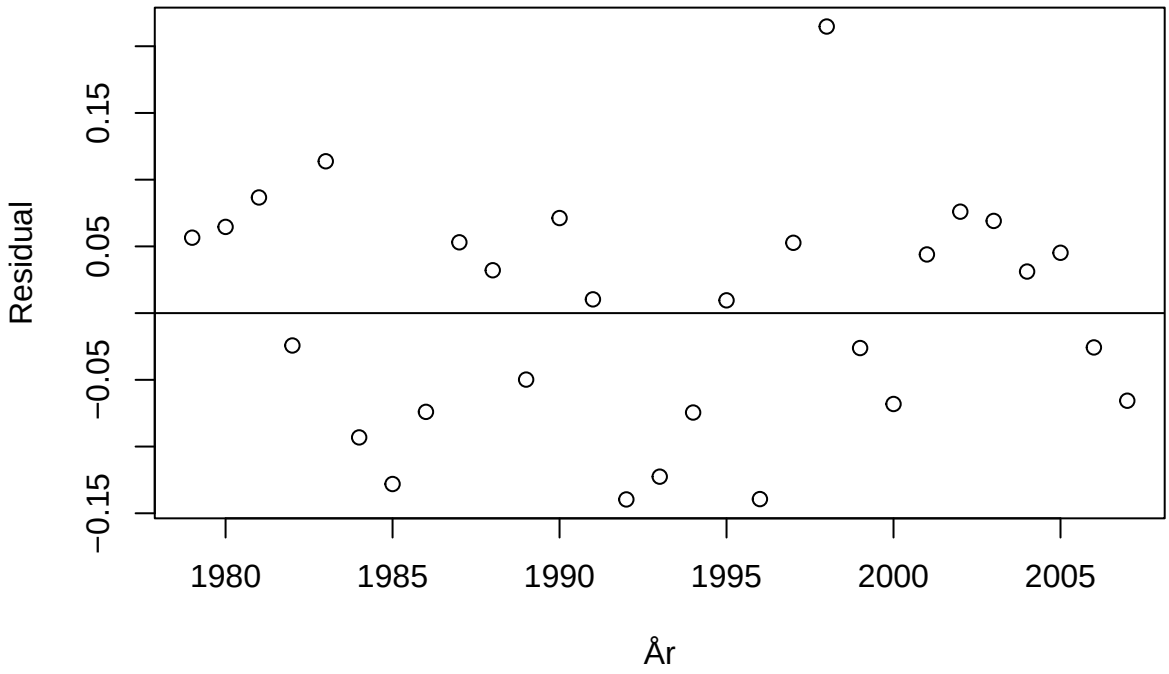

Figure 12: Residualplot över medeltemperatur och år under perioden 1979-2007

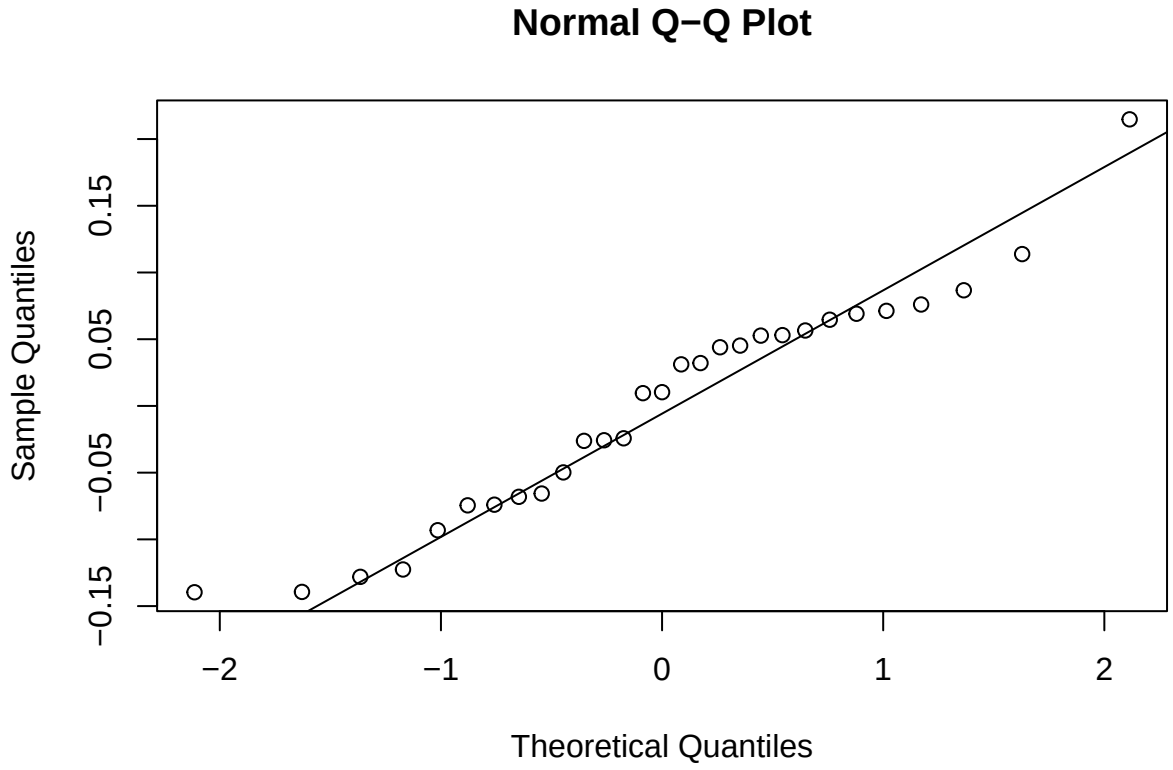

Figure 13: Normalfördelningsplot över residualen för medeltemperatur och år under perioden 1979-2007

```
## year        -0.0003018  0.0010281  -0.294    0.770
##
## Residual standard error: 0.1049 on 48 degrees of freedom
## Multiple R-squared:  0.001792,   Adjusted R-squared:  -0.019
## F-statistic: 0.08615 on 1 and 48 DF,  p-value: 0.7704
```

```
summary(lm_Model_2)
```

```
##
## Call:
## lm(formula = globe ~ year, data = df_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21349 -0.07483 -0.01996  0.09440  0.21462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0009988  2.8407046   0.352    0.727
## year        -0.0005744  0.0014571  -0.394    0.696
##
## Residual standard error: 0.1064 on 38 degrees of freedom
## Multiple R-squared:  0.004073,   Adjusted R-squared:  -0.02214
## F-statistic: 0.1554 on 1 and 38 DF,  p-value: 0.6956
```

```
summary(lm_Model_3)
```

```
##
```

```
## Call:
## lm(formula = globe ~ year, data = df_3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13962 -0.06810  0.01032  0.05654  0.21477
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.525836   3.852162  -8.703 2.56e-09 ***
## year          0.016935   0.001933   8.762 2.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08708 on 27 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7302
## F-statistic: 76.77 on 1 and 27 DF,  p-value: 2.233e-09
```

Tidsperioden 1880-1929 ser vi i diagram 7 inte uppfyller kraven för linjär regression eftersom residualen är längre ifrån normalfördelningen än den hela tidsperioden från uppgift 1.

För tisdperioden 1930-1969 ser vi i diagram 9 att data inte heller uppfyller kraven för linjär regression bättre än perioden från uppgift 1 eftersom residualernas varians är för avvikande, alltså att variansen för residualerna inte är konstant.

För tidsperioden från 1970-2007 ser vi i diagram 11, 12 och 13 att förutsättningarna för att uppfylla kraven för linjär regression är uppfyllda bättre än den hela tidsperioden från uppgift 1.

I figur 5 är punktskattningarna $\alpha = 0.2026480$ och $\beta == -0.0003018$. Där β är lutningskoefficienten och det som testas är den skattadelinjen.

I figur 8 är punktskattningarna $\alpha = 1.0009988$ och $\beta = -0.0005744$.

I figur 11 är punktskattningarna $\alpha = -33.525836$ och $\beta = 0.016935$.

Nu ska vi testa om påståendet ““Växthuseffekten är inget problem, det finns ingen bevisad trend mot ett varmare klimat” är korrekt. Vi behöver alltså bara undersöka om medeltemperaturen ökar eller ej ökar under perioden 1979-2007, vilket innebär att vi vill ha ett ensidigt konfidensintervall.

Kopplingen mellan oljedirektörens påstående och vår modell över den sökta perioden är frågan om globala medeltemperaturen har ökat linjärt eller om den inte har det.

Nollhypotesen är att globala medeltemperaturen inte ökar dvs. $H_0 : \beta = 0$.

Alternativhypotesen är att globala medeltemperaturen ökar dvs. $H_1 : \beta > 0$.

Eftersom Pearsons korrelationstest är mer lämplig för slumpvariabler och i detta fal är “år” inte en slumpvariabel blir Spearmans korrelationstest mer lämplig eftersom Spearman kollar ordningssambandet mellan den förklarande variabeln och relations variabeln.

```
cor.test(df_3$year, df_3$globe, alternative = "greater", method = "pearson", conf.level = 0.05)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_3$year and df_3$globe
## t = 8.762, df = 27, p-value = 1.116e-09
## alternative hypothesis: true correlation is greater than 0
## 5 percent confidence interval:
##  0.9241017 1.0000000
```

```
## sample estimates:
##       cor
## 0.8601257
```

Korrelationstestet tyder på en linjär ökning i medeltemperatur under perioden 1979-2007. Alltså kan vi bevisa alternativhypotesen och förkasta nollhypotesen med 5% signifikansnivå och därmed är direktörens påstående motbevisad.

Uppgift 3

En grupp fysiker har en hypotes att norra halvklotet påverkar klimatet på södra halvklotet med ett års fördröjning. Denna hypotes kommer undersökas genom att beräkna korrelationskoefficinten mellan medeltemperaturen på norra halvklotet ett visst år och medeltemperaturen på södra året efter. Den uppskattade korrelationskoefficienten blir 0.822 och p-värdet blir lika med noll.

```
ggplot(data = df, aes(x = year)) +
  geom_point(aes(y = nh)) +
  geom_point(aes(y = shnext, color = "red")) +
  geom_smooth(aes(y = nh)) +
  geom_smooth(aes(y = shnext, color = "red"))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
cor.test(df$nh, df$shnext, alternative = "two.sided", method = "pearson", conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$nh and df$shnext
## t = 17.939, df = 155, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7630897 0.8666464
## sample estimates:
##       cor
## 0.8215311
```

Vi ser från vårt test att fysikerna har räknat rätt med både p-värdet och korrelationskoefficienten.

Den höga korrelationen påverkas mer troligtvis av växthuseffekten, som både påverkar den norra och det södra halvklotets medeltemperatur. Slutsatsen som kan dras blir att temperaturen på det norra halvklotet inte skulle påverka medeltemperaturen på södra halvklotet. Båda halvkloten påverkas av samma effekt, växthuseffekten.

```
df_next <- subset(df, year <= 2007 & year >= 1982)

cor.test(df_next$nh, df_next$shnext, alternative = "two.sided", method = "pearson", conf.level = 0.90)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_next$nh and df_next$shnext
## t = 4.2392, df = 23, p-value = 0.0003102
```

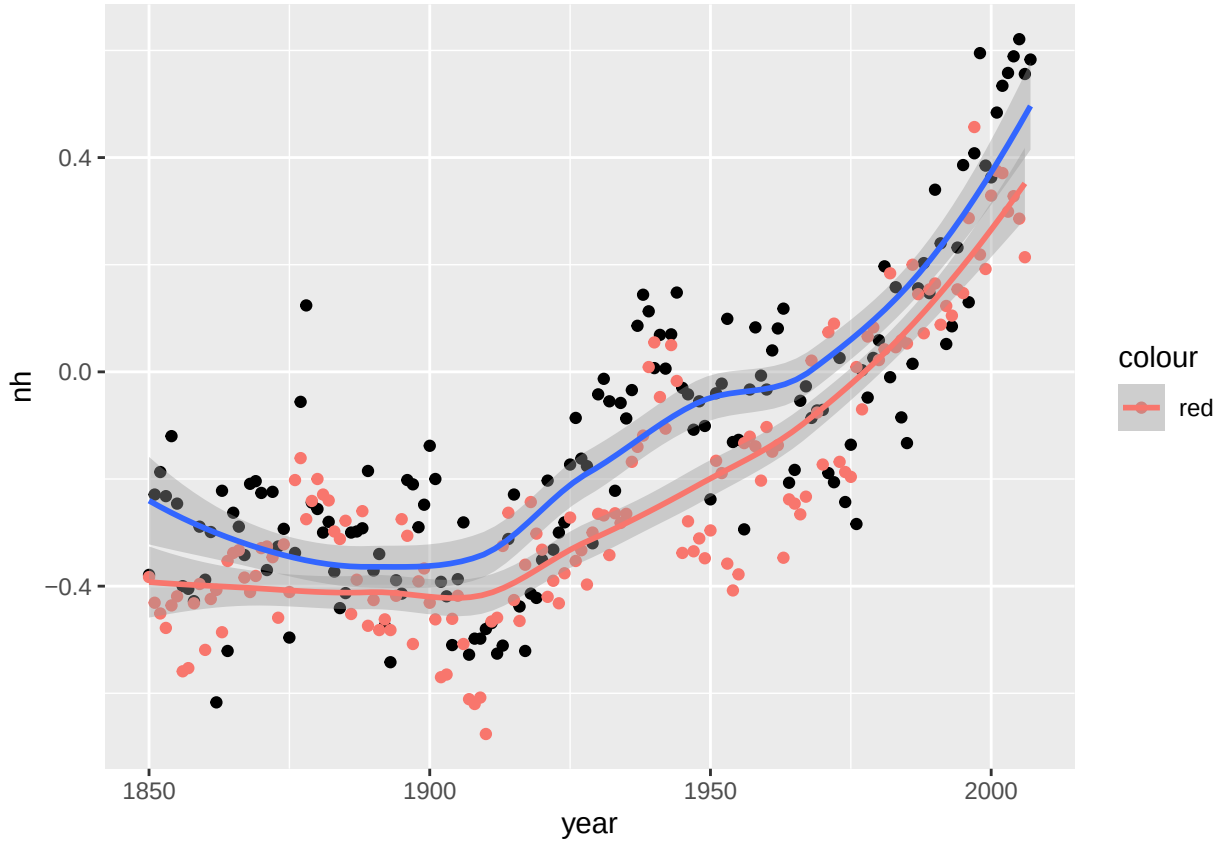

Figure 14: Scatterplott över temperaturen på norra halvklotet respektive södra halvklotet ett år senare, södra halvklotets mätvärden i rött

```
## alternative hypothesis: true correlation is not equal to 0
## 90 percent confidence interval:
##  0.4187639 0.8169442
## sample estimates:
##       cor
## 0.6622867
```

```
cor.test(df_next$nh, df_next$shnext, alternative = "two.sided", method = "spearman", conf.level = 0.90)
```

```
## Warning in cor.test.default(df_next$nh, df_next$shnext, alternative =
## "two.sided", : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_next$nh and df_next$shnext
## S = 909.67, p-value = 0.000435
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.650125
```

```
plot(df_next$nh, df_next$shnext, xlab="Norra halvklotets medeltemperatur", ylab="Södra halvklotets medel
```

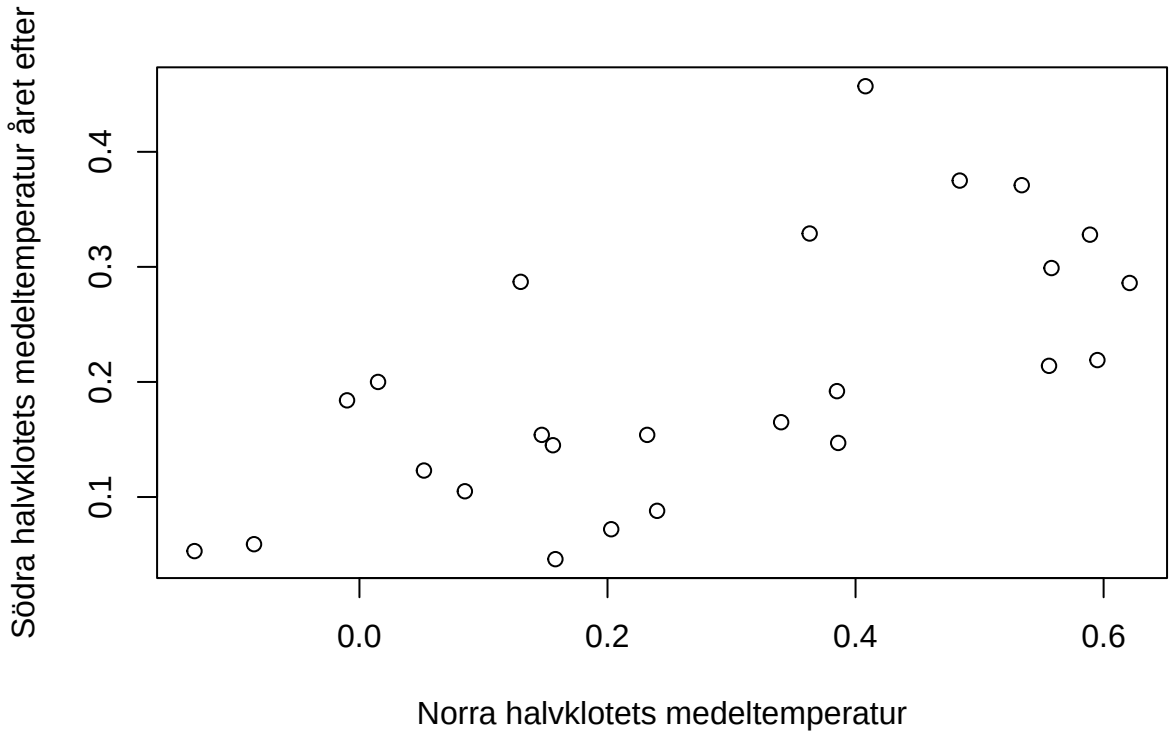

Figure 15: Scatterplot över medeltemperatur mellan halvkloten