Statistik för Biologer
F6: Linjär Regression

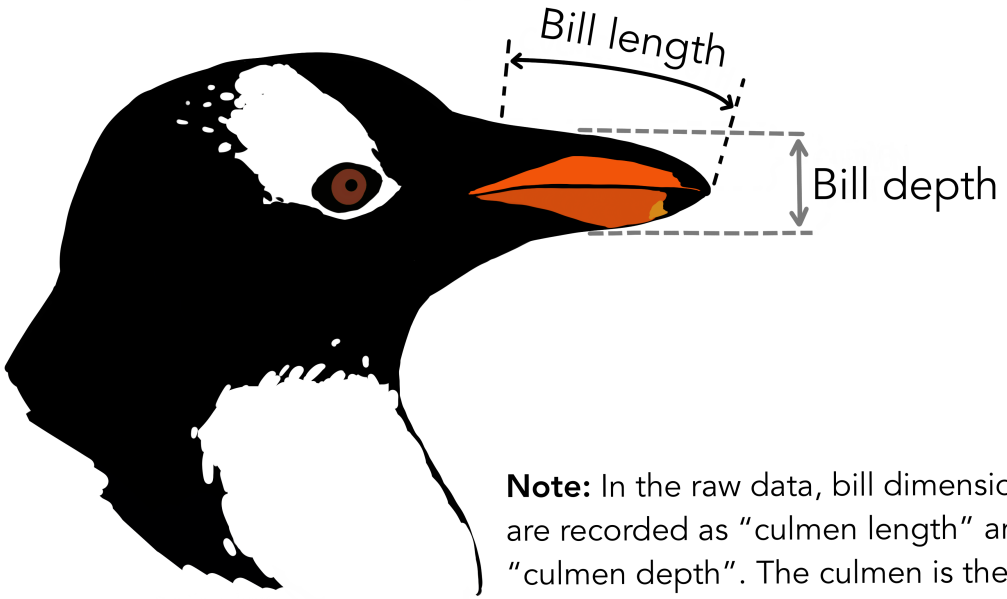
Shaobo Jin

Matematiska institutionen

Att Mäta Pingviner

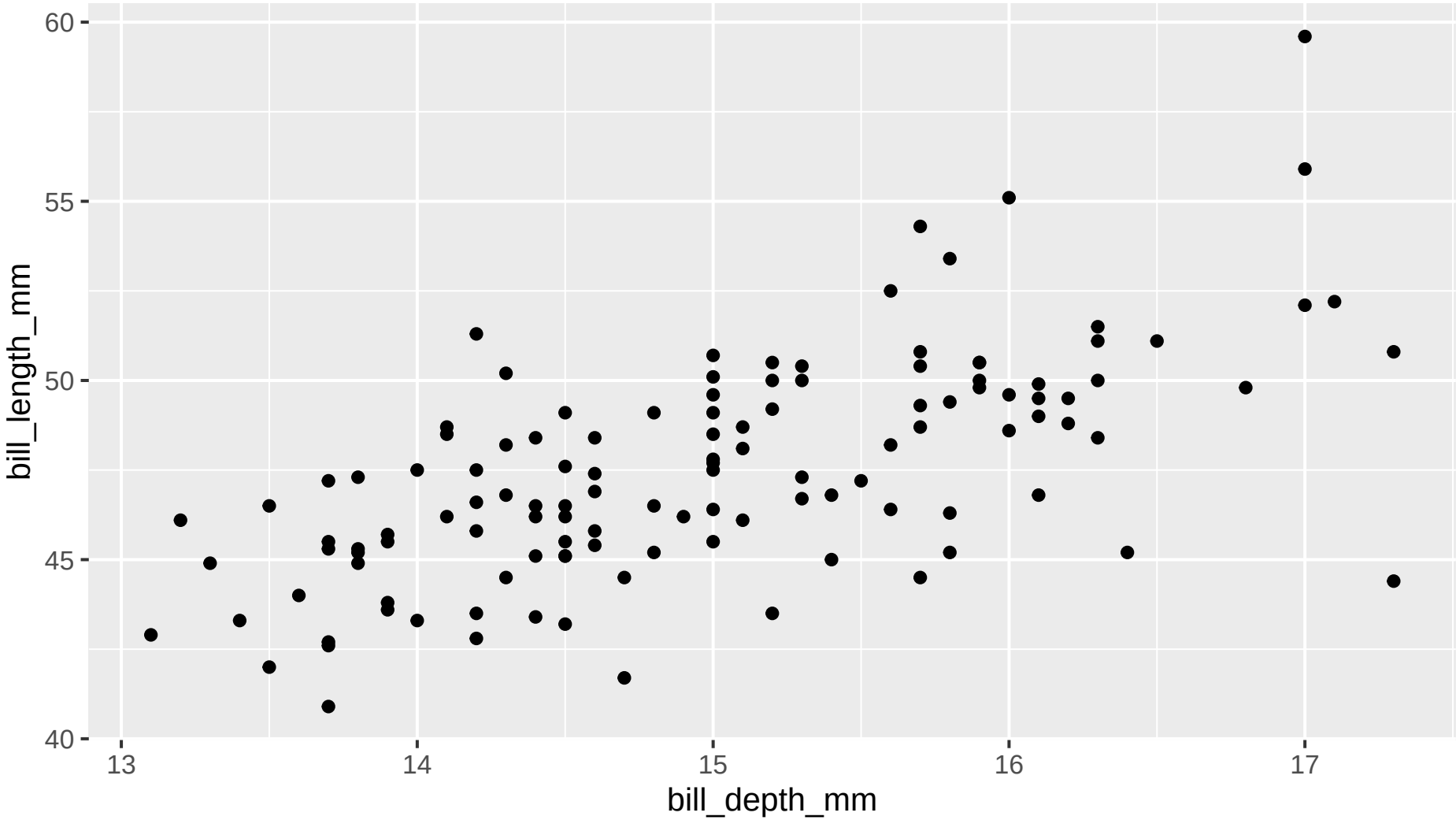
Det finns tre arter av pingviner. Forskarna har mätt bland annat

1. Näbbens längd (mm)
2. Näbbens djup (mm)
3. Vingens längd (mm)
4. Vikt (g)
5. Kön (hona/hane)
6. Art (tre arter)



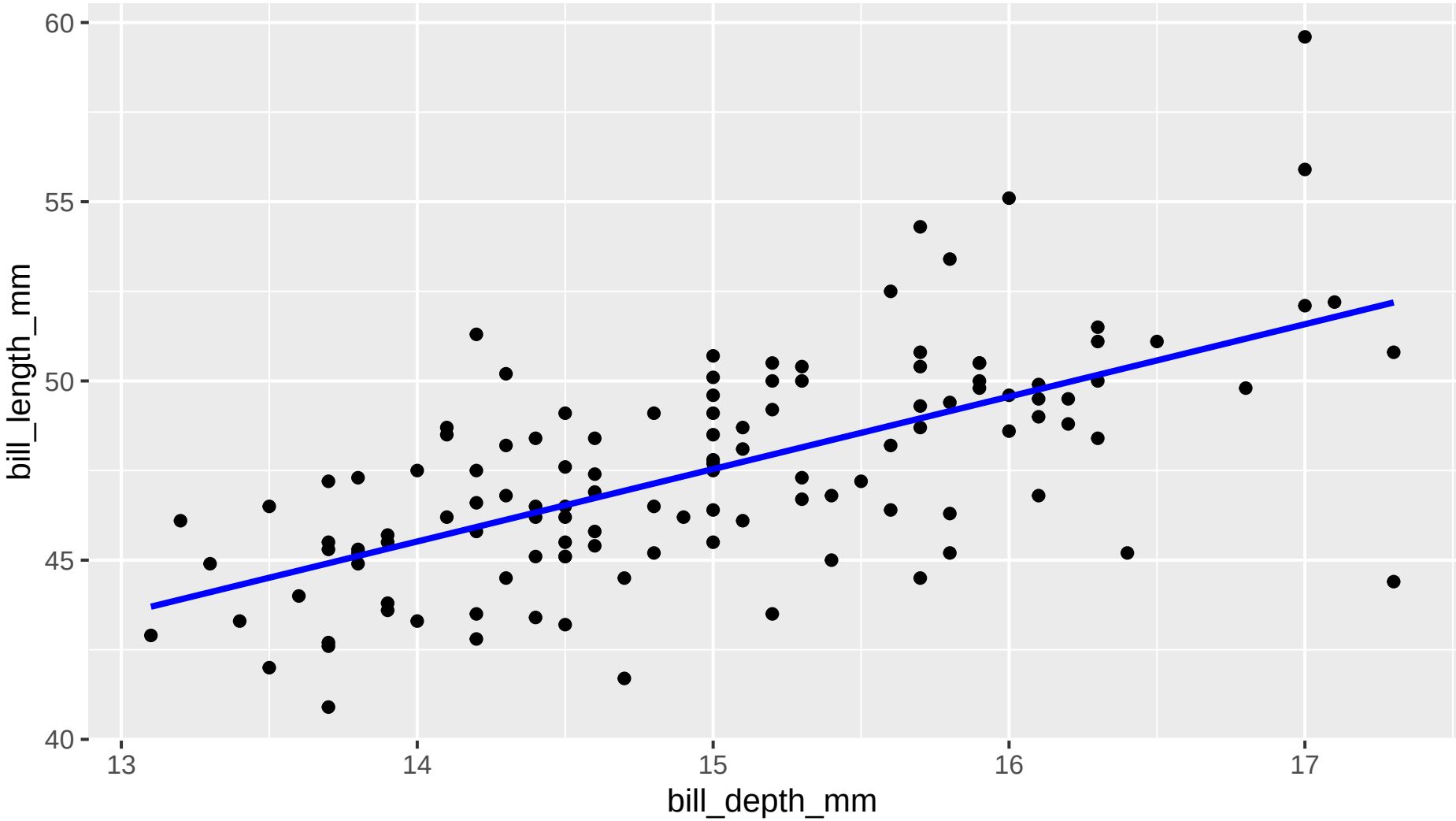
Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Spridningsdiagram: Gentoo



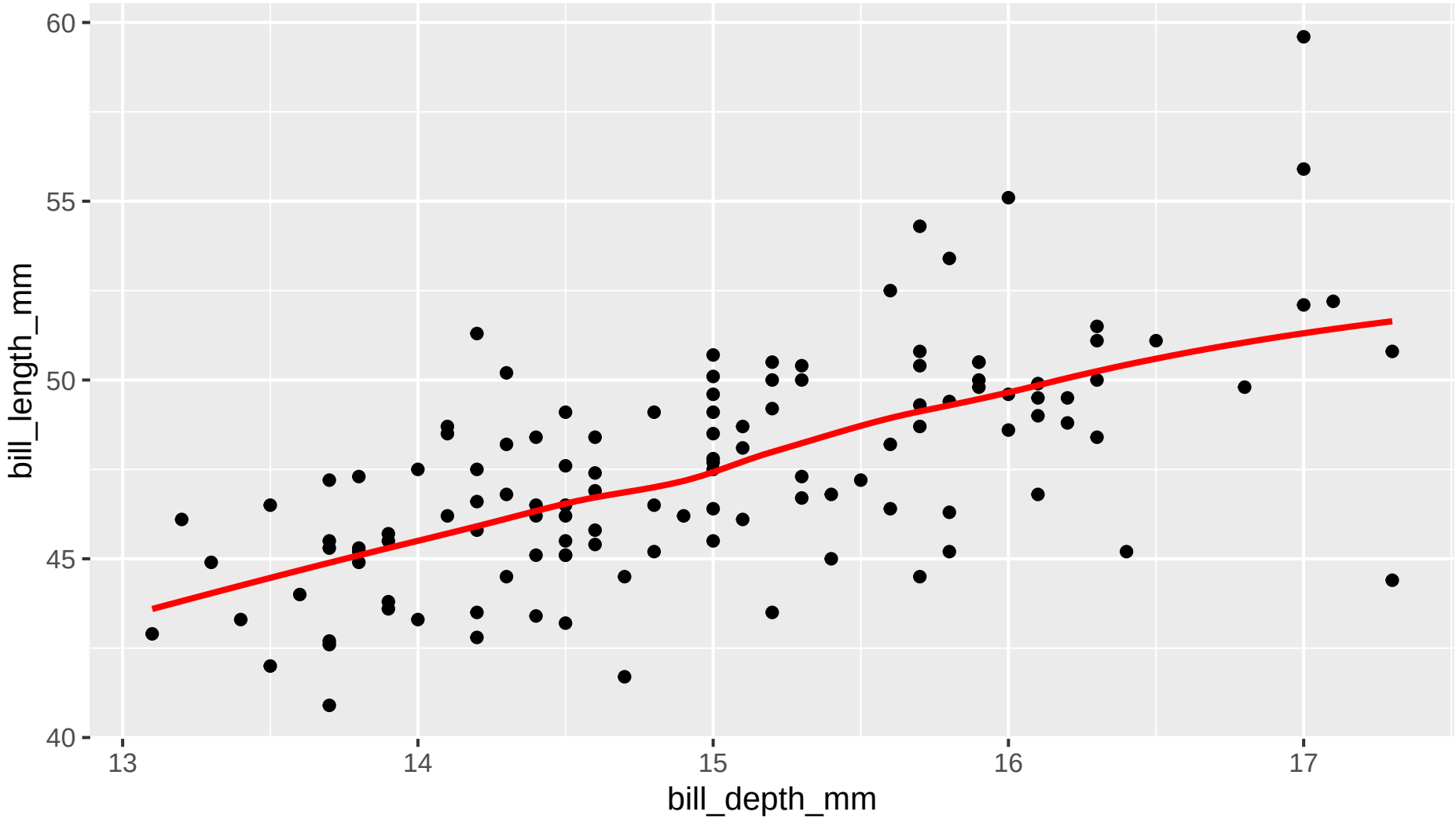
Spridningsdiagram: Gentoo

Vi vill hitta en funktion som beskriver sambandet mellan näbbens längd och näbbens djup. Men vilken funktion beskriver sambandet?



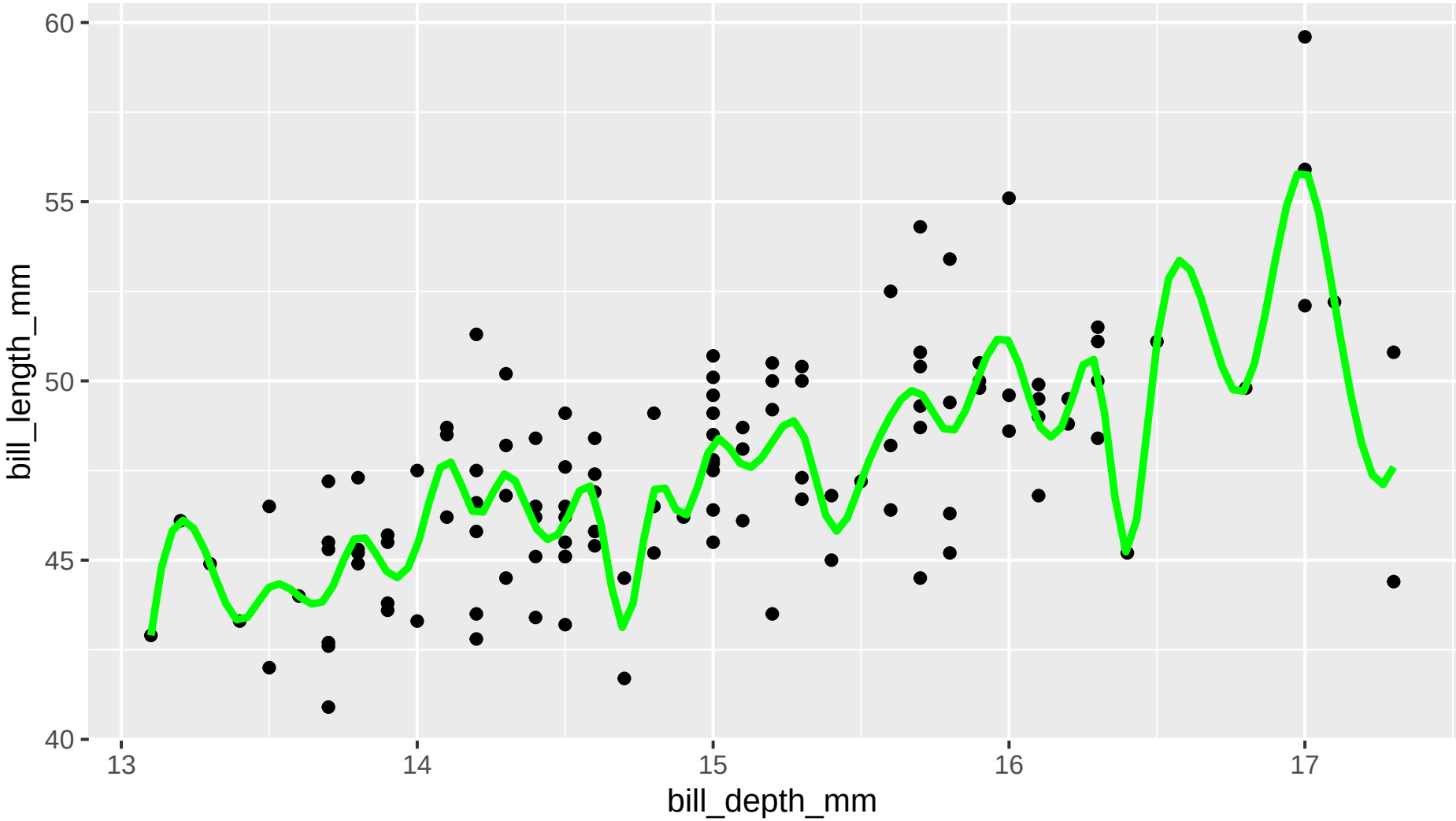
Spridningsdiagram: Gentoo

Vi vill hitta en funktion som beskriver sambandet mellan näbbens längd och näbbens djup. Men vilken funktion beskriver sambandet?



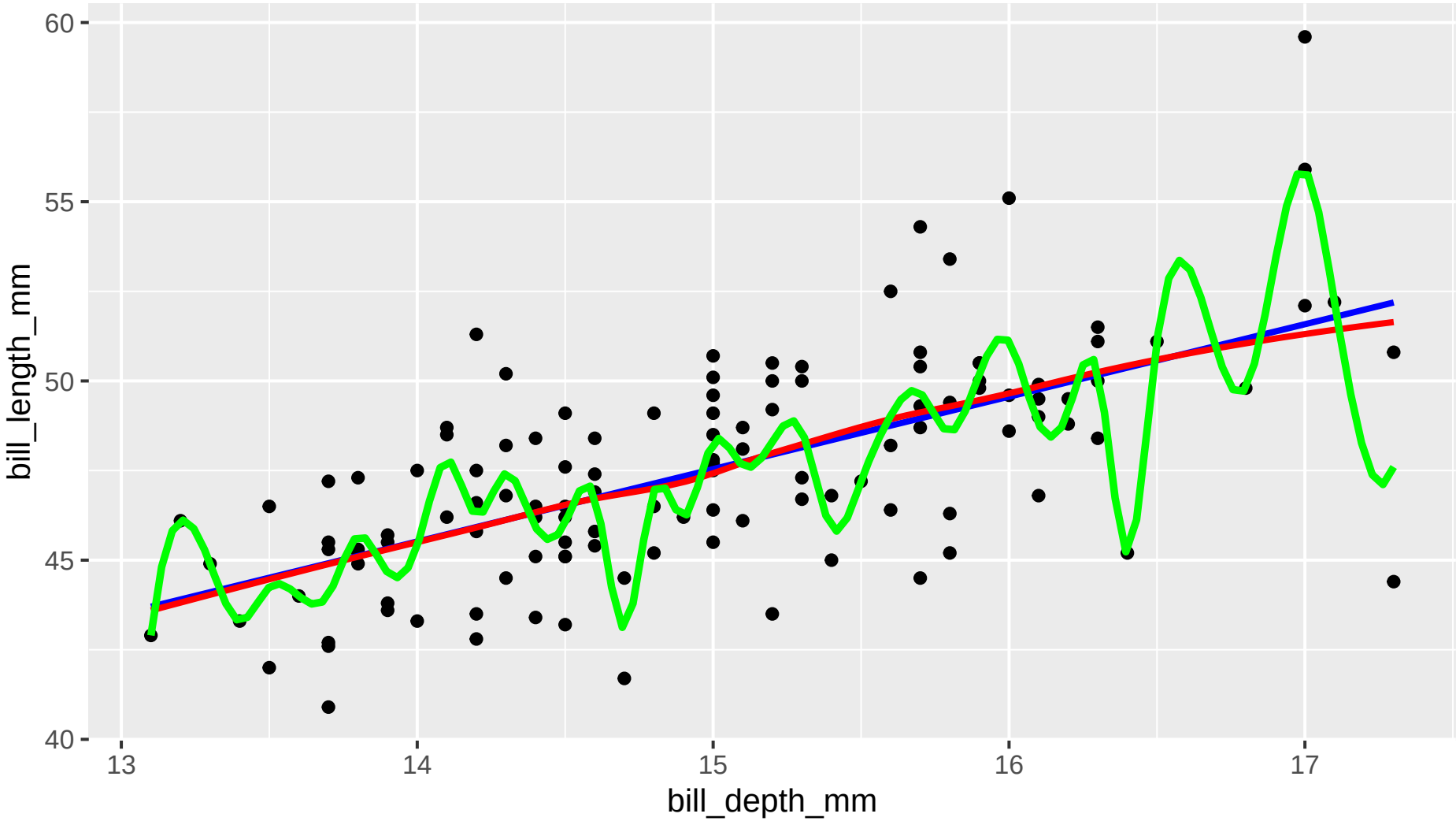
Spridningsdiagram: Gentoo

Vi vill hitta en funktion som beskriver sambandet mellan näbbens längd och näbbens djup. Men vilken funktion beskriver sambandet?



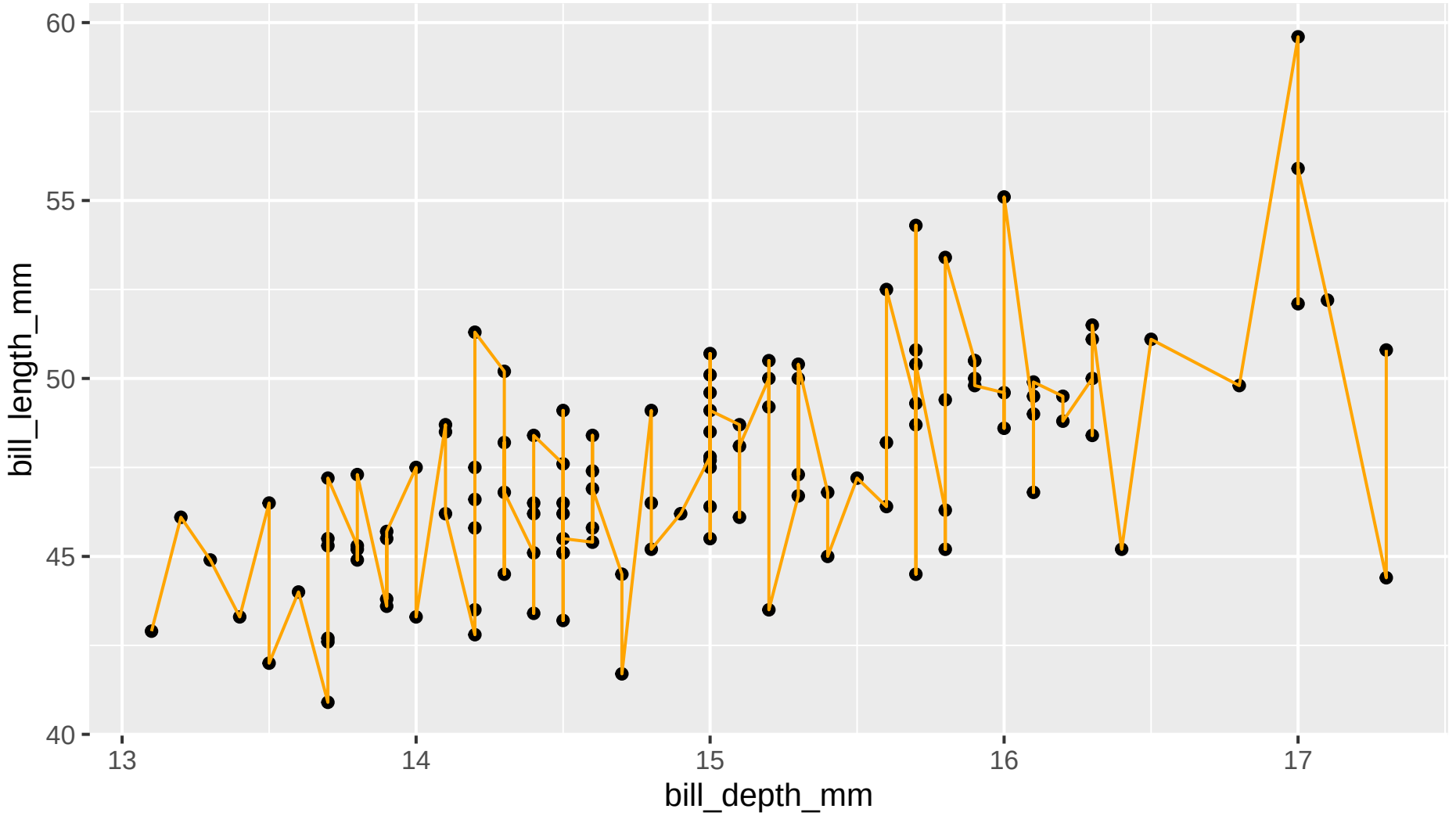
Spridningsdiagram: Gentoo

Vi vill hitta en funktion som beskriver sambandet mellan näbbens längd och näbbens djup. Men vilken funktion beskriver sambandet?



Spridningsdiagram: Gentoo

Vi kan hitta en funktion som passar data perfekt:



Vi måste begränsa urvalet av linjer!

Praktiska önskemål:

- Sambandet ska om möjligt beskrivas av en enkel funktion som går att tolka
- Funktionen ska inte passa "för bra" med våra mätdata - vi vill undvika **överanpassning**
- Det vore bra om vi har en sorts funktion som går att använda i många olika tillämpningar

Förslag: börja med **linjära funktioner** på formen $y = kx + m$:

- Har enkel tolkning
- Många samband är approximativt linjära
- Många samband kan transformeras till linjära funktioner

Modell

Linjär regression används för att studera hur en **responsvariabel** Y_i beror på en **förklarande variabel** x_i .

Linjär modell:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- Y_i är något vi mäter (responsvariabeln eller den beroende variabeln)
- x_i är bestämd av oss eller mätt utan (nämnvärt) mätfel (den oberoende variabeln eller förklarande variabeln)
- ϵ_i är en slumpavvikelse (mätfel): avvikelser från linjen orsakade av andra faktorer än x

Vi har gjort n mätningar (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n) och vill använda dessa för att bestämma värdena på α och β.

Användningsområden

Beskriva samband

- Beror tillväxthastighet hos cyanobakterier på temperaturen?
- Hur mycket förändras tillväxthastigheten om temperaturen ändras två grader?

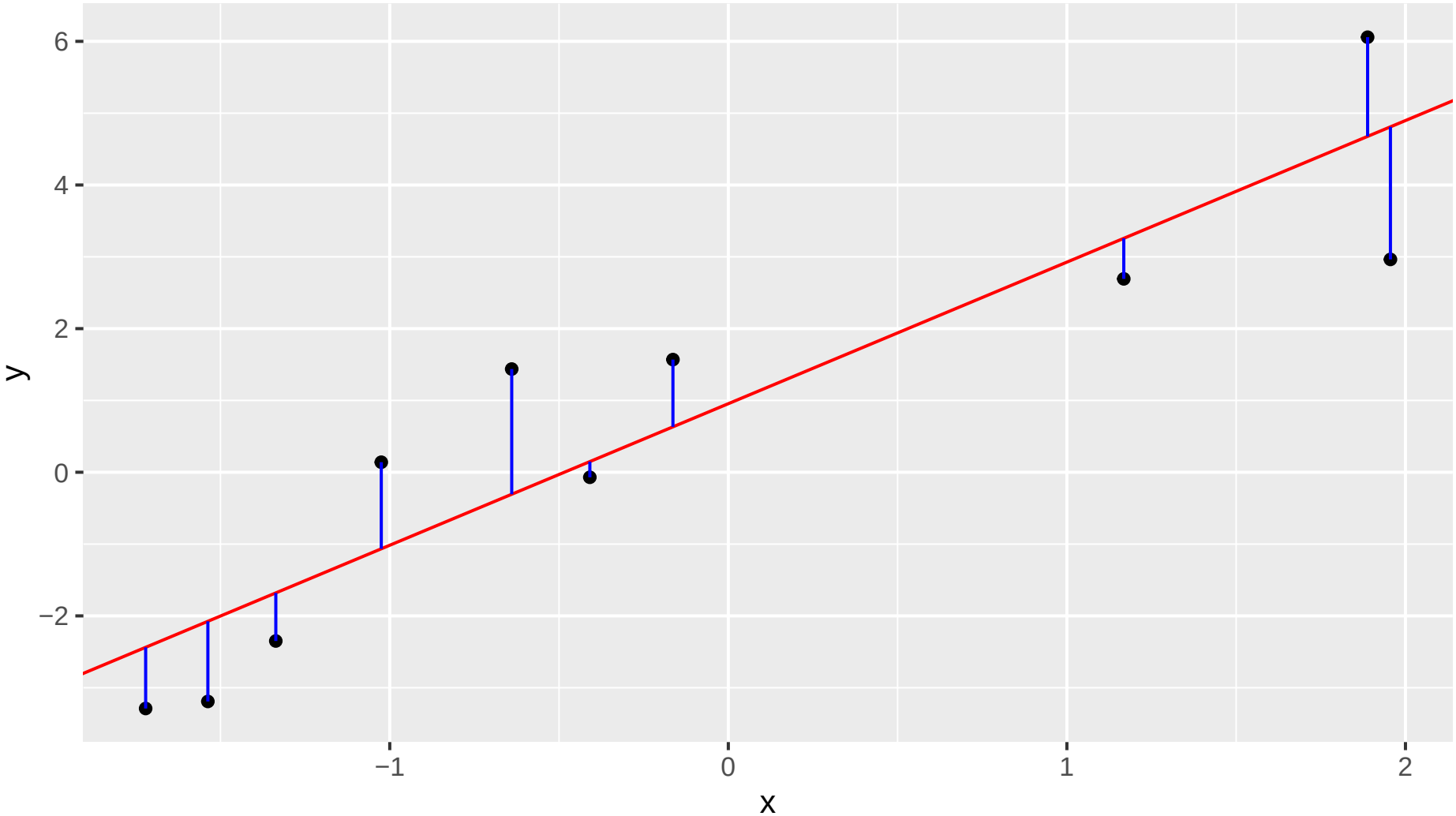
Göra prediktion: modellen låter oss göra prediktioner om värdet på y, givet nya värden på x. Om både α och β är kända blir prediktionen $\alpha + \beta x$.

- Vad blir tillväxthastigheten för cyanobakterier vid temperaturen x?

Maskininlärning och artificiell intelligens

Avstånd till linjen

För varje möjlig linje kan vi mäta avståndet mellan linjen och våra datapunkter, $(\alpha + \beta x_i) - y_i$:



Vilken linje passar data bäst?

Vi skattar parametrarna α och β (anpassar modellen) med **minsta kvadrat-metoden**:

- $(\alpha + \beta x_i) - y_i$ är avståndet mellan linjen (prediktionen) och y_i.
- $[(\alpha + \beta x_i) - y_i]^2$ är kvadratavståndet mellan linjen (prediktionen) och y_i.
- Hitta de värden på α och β så att $\sum_{i=1}^{n} [(\alpha + \beta x_i) - y_i]^2$ är så liten som möjligt!

Går att ta fram formler för skattningarna med hjälp av linjär algebra, men i praktiken används nästan alltid ett statistikprogram för beräkningarna.

R-kod för modellanpassning

I allmänhet

```
lm(responsvariabeln ~ foerklarande variablerna, # formel
   data = datasnamn) # data
```

I vårt exempel:

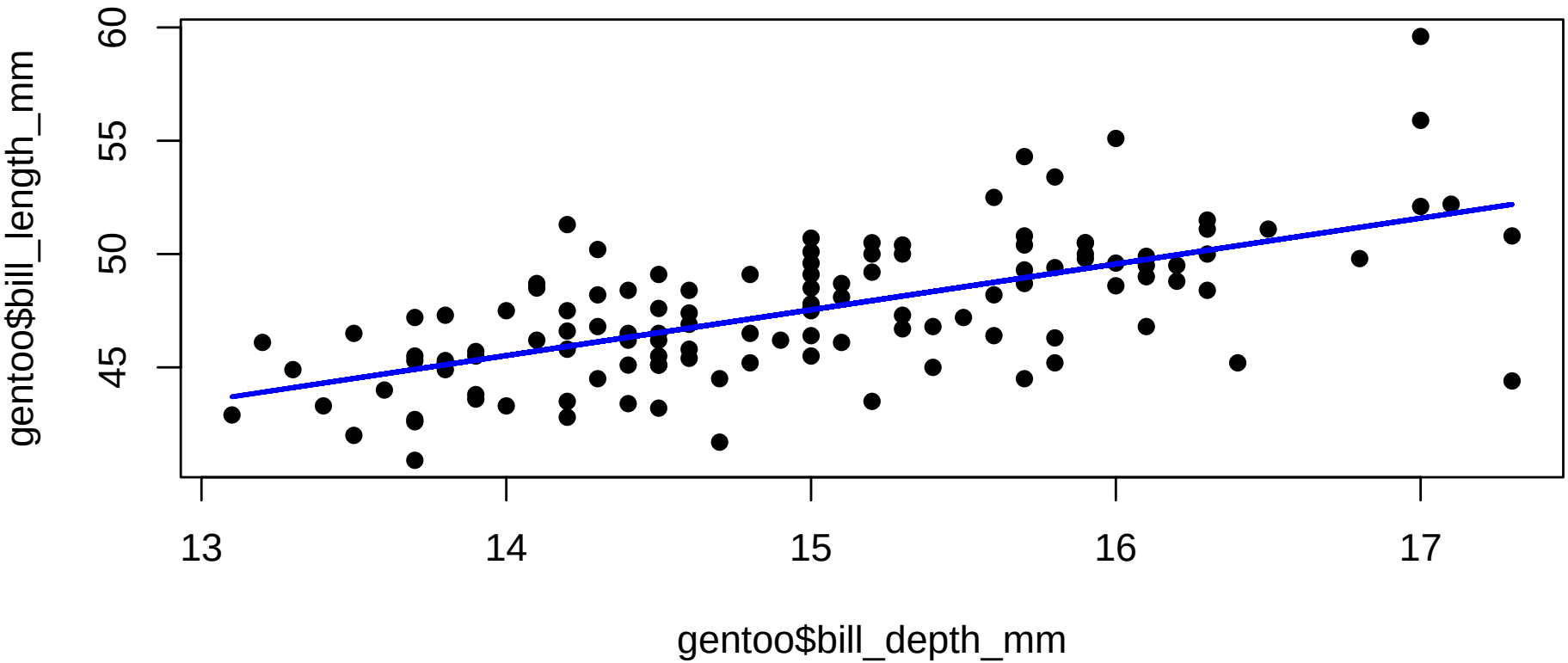
```
LM <- lm(bill_length_mm ~ bill_depth_mm, data = gentoo)
LM

##
## Call:
## lm(formula = bill_length_mm ~ bill_depth_mm, data = gentoo)
##
## Coefficients:
##   (Intercept)  bill_depth_mm
##        17.230          2.021
```

Anpassad modell

```
summary(LM) # Ta fram den viktigaste information

##
## Call:
## lm(formula = bill_length_mm ~ bill_depth_mm, data = gentoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7888 -1.4097  0.1361  1.3882  8.0174
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.2295     3.2818   5.250 6.60e-07 ***
## bill_depth_mm   2.0208     0.2186   9.245 1.02e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.369 on 121 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4139,Adjusted R-squared:  0.4091
## F-statistic: 85.46 on 1 and 121 DF,  p-value: 1.016e-15
```

Visualisering av anpassad linje

```
prediktion <- predict(LM)
plot(gentoo$bill_depth_mm, gentoo$bill_length_mm, pch = 16)
lines(na.omit(gentoo$bill_depth_mm), prediktion, col = "blue")
```



Konfidensintervall

Konfidensintervall för α och β:

```
confint(LM, level = 0.95) # alpha = 0.05
```

```
##                   2.5 %    97.5 %
## (Intercept)   10.732217 23.726785
## bill_depth_mm  1.588018  2.453518
```

p-värden

I summary-tabellen visas p-värden för olika hypoteserna.

1. Pr(>|t|) för interceptet α:

$$H_0 : \alpha = 0 \text{ (Då } x = 0 \text{ är } y = 0)$$
$$H_1 : \alpha \neq 0 \text{ (Då } x = 0 \text{ är } y \neq 0)$$

2. Pr(>|t|) för lutningen β:

$$H_0 : \beta = 0 \ (x \text{ påverkar inte } y)$$
$$H_1 : \beta \neq 0 \ (x \text{ påverkar } y)$$

   Vi är oftast mest intresserade av resultatet för β.

3. F-statistic, p-value:

$$H_0 : \text{alla } \beta = 0$$
$$H_1 : \text{någon/några } \beta \neq 0$$

F-statistic och p-värdet

```
summary( lm(bill_length_mm ~ bill_depth_mm + flipper_length_mm, data = gentoo) )

##
## Call:
## lm(formula = bill_length_mm ~ bill_depth_mm + flipper_length_mm,
##     data = gentoo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.022 -1.415  0.016  1.290  7.353
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -11.63177    7.07676  -1.644 0.102865
## bill_depth_mm       1.10533    0.28674   3.855 0.000188 ***
## flipper_length_mm   0.19604    0.04339   4.518 1.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.199 on 120 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4992,Adjusted R-squared:  0.4908
## F-statistic:  59.8 on 2 and 120 DF,  p-value: < 2.2e-16
```

Förklaringsgrad R^2

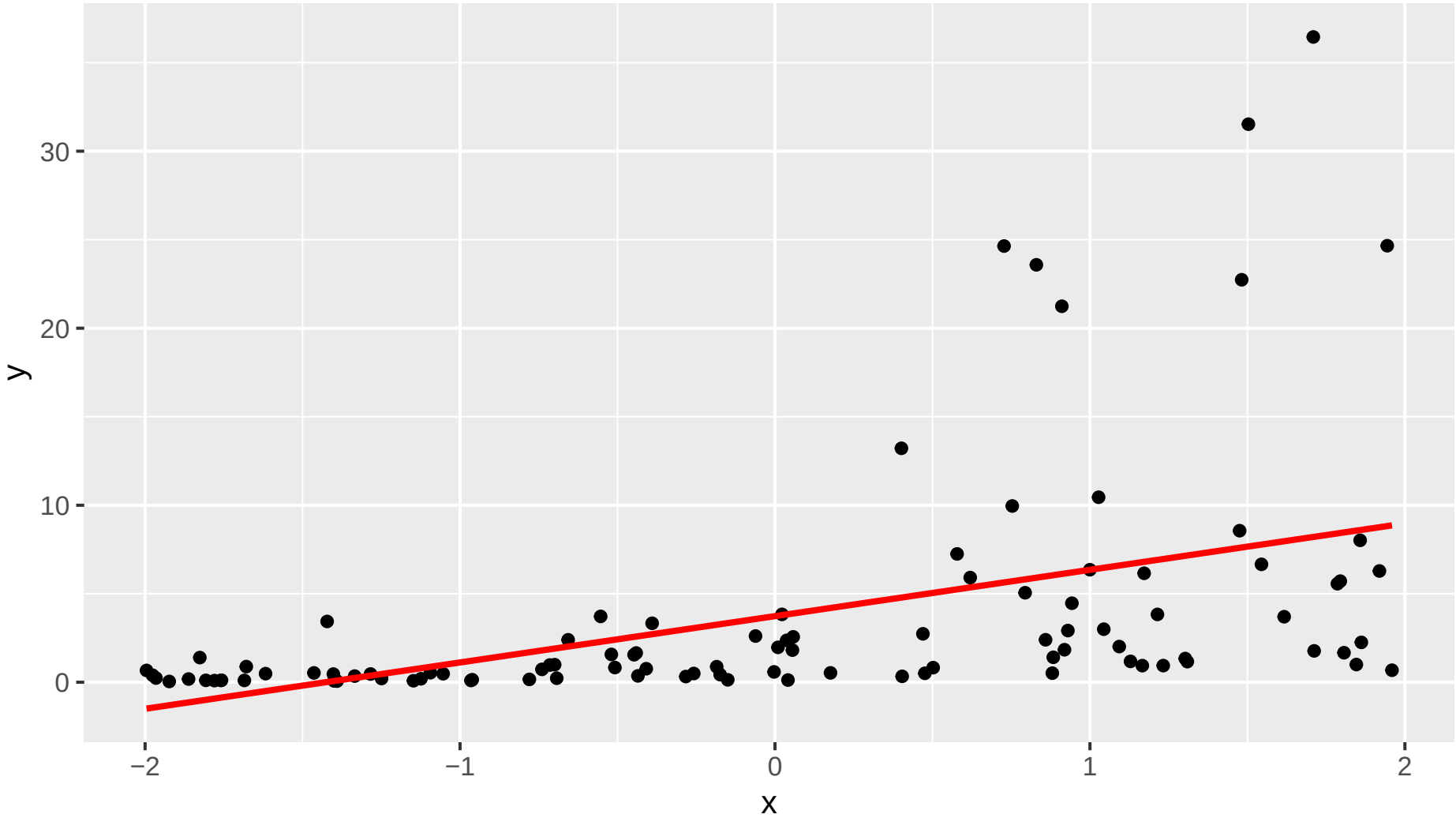
I summary-tabellen visas också måttet **Multiple R-squared**.
Förklaringsgraden R^2 anger hur mycket av variationen i y som förklaras av x. Om vi bara har en förklarande variabel, $R^2 = \rho^2$.

```
cor(gentoo$bill_length_mm, gentoo$flipper_length_mm,
    use = "complete.obs") ^ 2

## [1] 0.4371354
```

```
summary(LM)$r.squared

## [1] 0.4139429
```

Modellförutsättningar 1: Linearitet

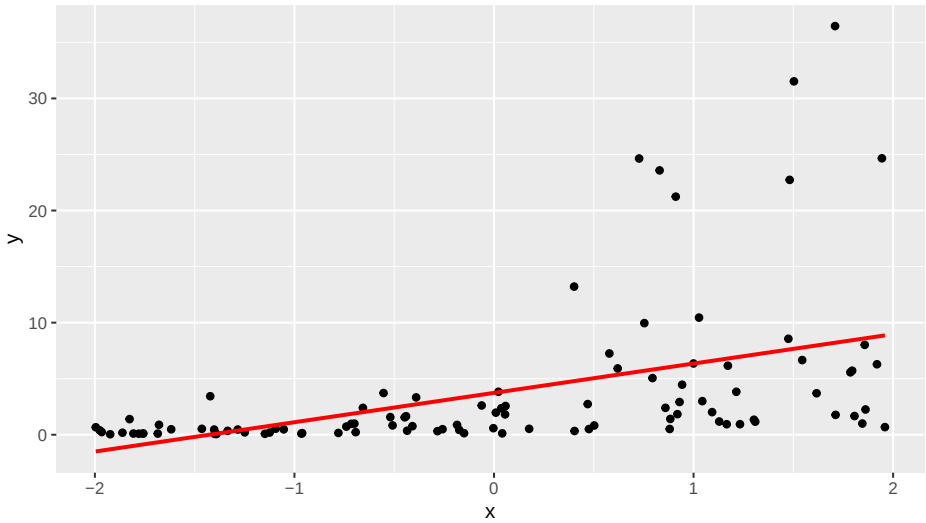
Om sambandet inte beskrivs tillräckligt bra av $y = \alpha + \beta x$ så kommer modellen inte fungera särskilt bra.

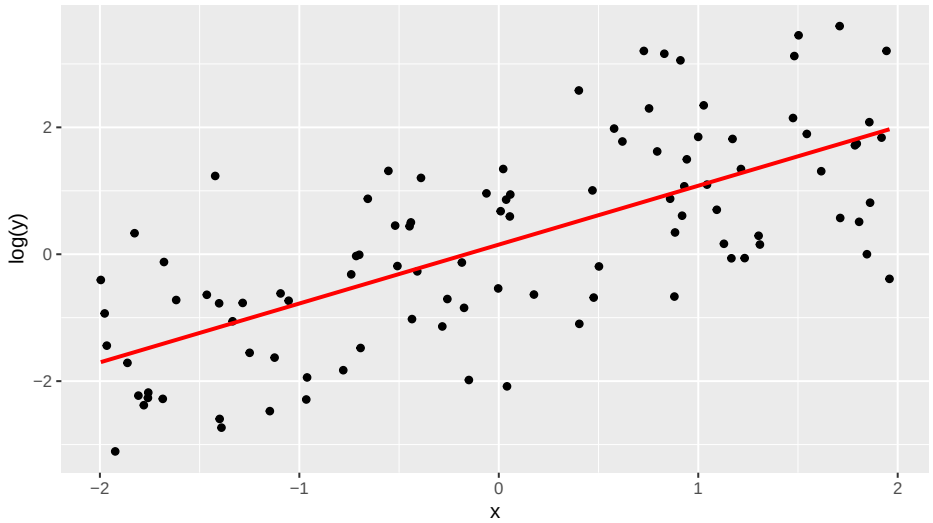


Modellförutsättningar 1: Linearitet

Om sambandet inte är linjärt kan vi ibland transformera y

- Logaritmering, $\log(y)$
- Kvadratrot $\sqrt{y}$
- Och så vidare. . .

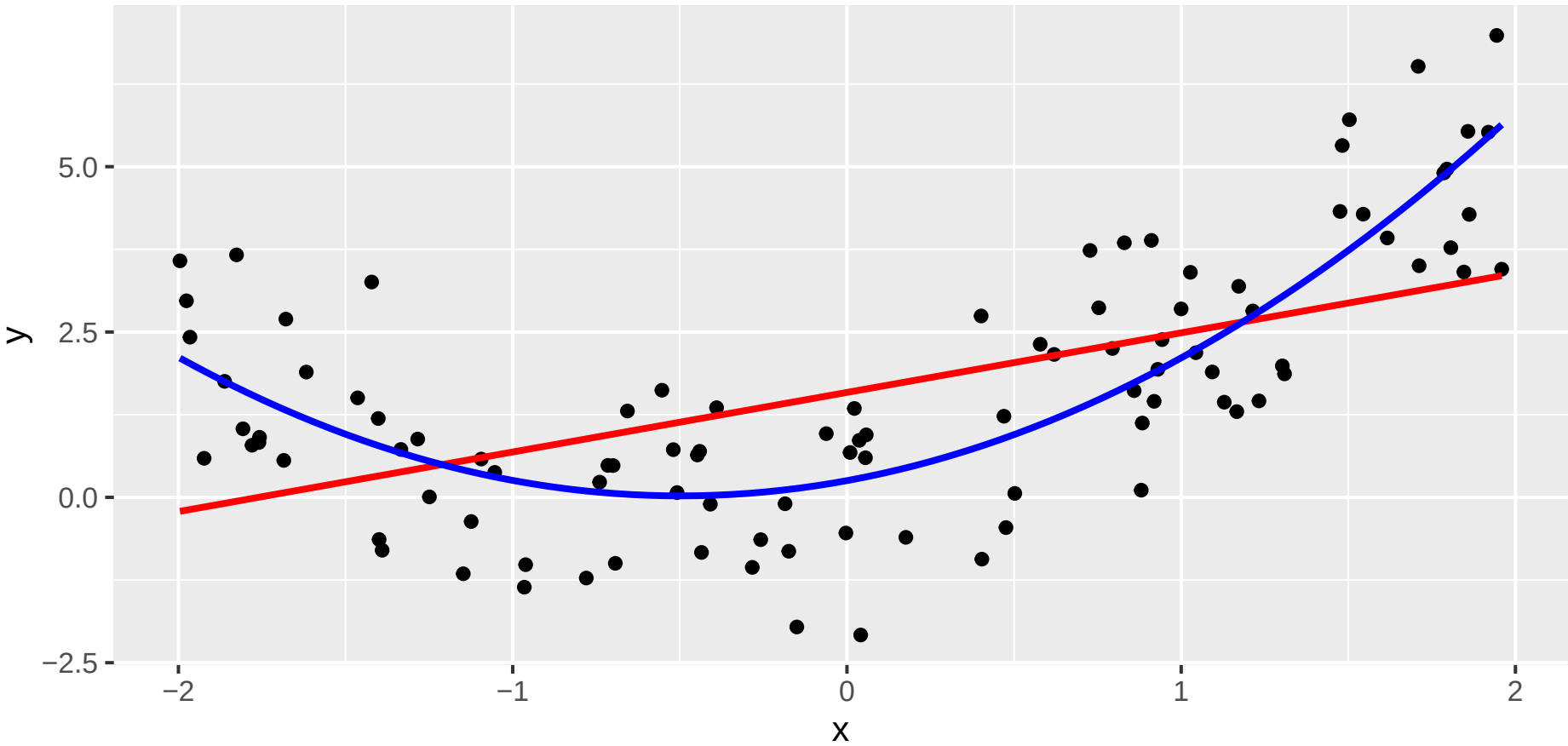




Modellförutsättningar 1: Linearitet

Om sambandet inte är linjärt kan vi t.ex. använda "icke-linjär" x-variabel. Vi kan stoppa in x^2 eller 2^x och så vidare i modellen.

```
lm(y ~ x + I(x ^ 2), data = Data) # den kvadratiska modellen
```



Modellförutsättningar 2: Lika varians

Variationen i y-variabeln ska vara lika stora för alla värden på x, t.ex.

$$V(\epsilon_1) = V(\epsilon_2) = \cdots = V(\epsilon_n).$$

Lika varians krävs för konfidensintervall och p-värden!

Vad kan vi göra om spridningen i y är olika för olika värden på x?

1. Prova att logaritmera y-variabeln. Leder ofta till "**variansstabilisering**"
2. Använd vissa sorters **bootstrap** för beräkning av p-värden och konfidensintervall

Modellförutsättningar 3: Normalfördelning

Slumpavvikelserna ϵ_i ska vara normalfördelade, vilket krävs för konfidensintervall och p-värden!

Vad kan vi göra om slumpavvikelserna ϵ_i inte är normalfördelade?

1. Prova att logaritmera y-variabeln. Leder ofta till "mer normalfördelade" slumpavvikelser.
    1. Eller ▸ Box-Cox transformation
2. Använd **bootstrap** för beräkning av p-värden och konfidensintervall. Mer om det på Föreläsning 7.

Bootstrap

R-kod för att använda bootstrap för p-värden och konfidensintervall:

```
library(boot.pval)
LM <- lm(bill_length_mm ~ bill_depth_mm, data = gentoo[-120,])
boot_summary(LM)

##                Estimate Lower.bound Upper.bound p.value
## (Intercept)   17.229501   10.760529   23.924022       0
## bill_depth_mm  2.020768    1.583982    2.449103       0
```

Nonsenssamband och falska upptäckter

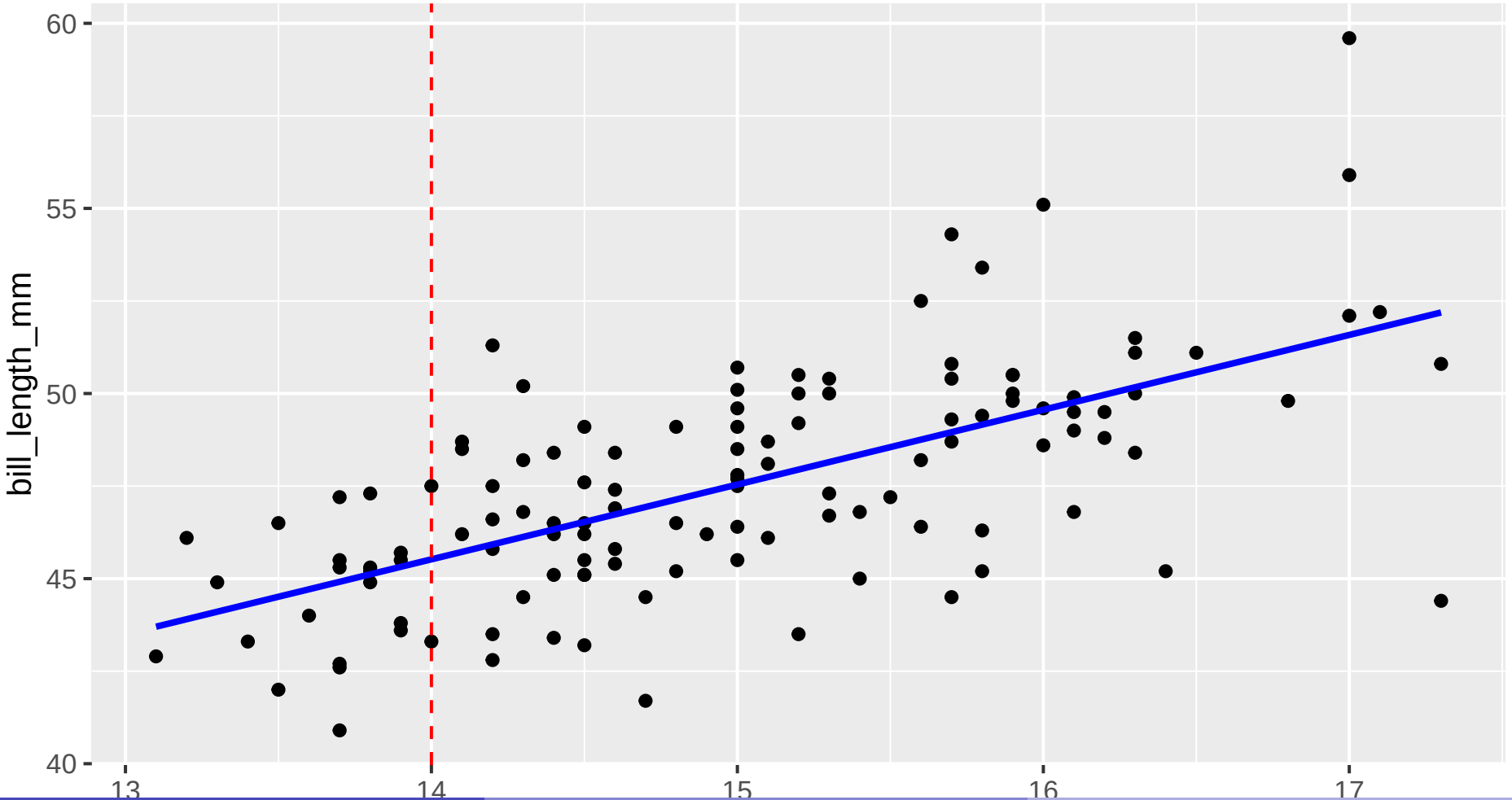
Regression är en beskrivning av samvariationen mellan två eller flera variabler. Korrelation och regression är inte samma sak som **kausalitet**: alla samband är inte orsakssamband. Exempel:

- Glassförsäljning har en hög korrelation med antalet drunkningsolyckor
- Korrelationen mellan antalet häckande storkar och antalet födda barn i Köpenhamn 1945-1960 är hög

Om vi kör en massa olika regressionsmodeller på våra data så kommer vi förr eller senare få signifikanta samband av ren slump. Undvik "datafiske" där du testar massor av olika samband i dina data. Låt forskningshypoteserna styra vilka analyser du gör.

Att använda modellen

Man vill ofta använda sin regressionmodell för att göra **prediktioner**.
Exempel: Vad är den förväntade näbbenslängden om näbbensdjupet är 14mm?



Att använda modellen

Vi använder våra anpassade koefficienter $\hat{\alpha} = 17.2295$ och $\hat{\beta} = 2.0208$ för prediktionen:

$$y = 17.2295 + 2.0208 \cdot 14 = 45.5207.$$

Beräkning med R

```
LM <- lm(bill_length_mm ~ bill_depth_mm, data = gentoo)
predict(LM, newdata = data.frame(bill_depth_mm = 14))
```

```
##        1
## 45.52025
```

Svaret avviker lite eftersom R använder fler decimaler för $\hat{\alpha}$ och $\hat{\beta}$.

Prediktionsintervall

Vi kan få ett **prediktionsintervall** för en ny observation skulle kunna hamna då $x = 14$mm.
Det tar dels hänsyn till osäkerheten i våra skattningar av α och β och dels hänsyn till hur stora slumpavvikelserna från linjen brukar vara:

```
predict(LM, newdata = data.frame(bill_depth_mm = 14),
        interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 45.52025 40.79198 50.24853
```

interval = "prediction" Är Viktigt!

Prediktionsintervall för en ny observation

```
predict(LM, newdata = data.frame(bill_depth_mm = 14),
        interval = "prediction", level = 0.95)

##        fit      lwr      upr
## 1 45.52025 40.79198 50.24853
```

Konfidensintervall för $\alpha + \beta x$ (det förväntade värdet)

```
predict(LM, newdata = data.frame(bill_depth_mm = 14),
        interval = "confidence", level = 0.95)

##        fit      lwr      upr
## 1 45.52025 44.92069 46.11981
```

Extrapolering

Man måste alltid vara lite försiktig med att extrapolera modellen till x-värden som ligger lång ifrån de data vi använt för att anpassa modellen.

- Sambandet kan ändra form utanför det x-område vi har studerat
- Skörden av en gröda ökar vid mer bevattning. Men när bevattningen blir för stor så minskar skörden igen.

Sammanfattning

1. Linjär regression används för att modellera samband av typen $y = kx + m$
2. Vi anpassar modellen utifrån parvisa mätningar av (x, y)
3. Förutsättningar:
   1. Linearitet
   2. Lika varians
   3. Normalfördelning