F12 – Regression

Måns Thulin

Uppsala universitet

thulin@math.uu.se

Statistik för ingenjörer • 28/2 2013

Dagens föreläsning

- Linjära regressionsmodeller
- Stokastisk modell
- Linjeanpassning och skattningar
- Konfidensintervall
- Prediktion och extrapolering

Repetition från F1: Beroendemått

Ofta mäter man två olika egenskaper för varje enhet (som i borrexemplet, där man mätte borrad längd och nötning).

Man har då två variabler x och y som finns registrerade parvis: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Beroendemått beskriver *samvariationen* mellan de två variablerna.

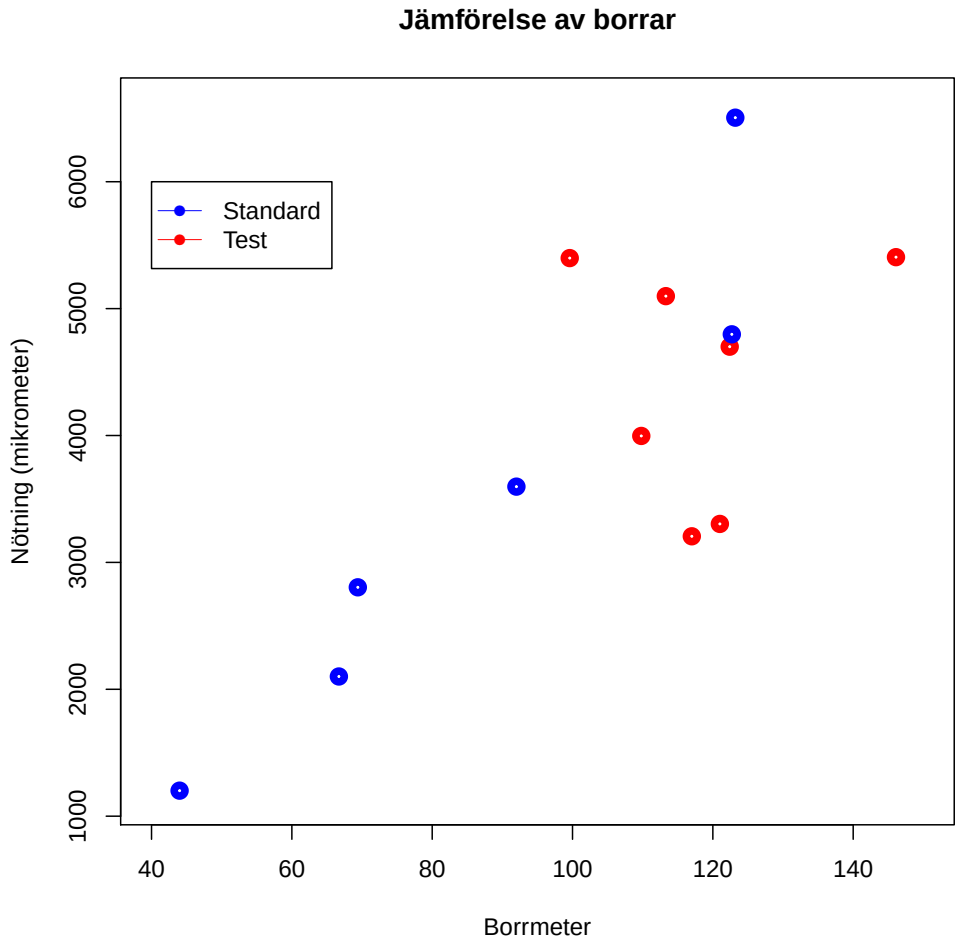
Korrelationskoefficienten:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

är en enhetslös storhet sådan att $-1 \leq r \leq 1$.

- Om $r = 1$ så ligger observationerna på en rät linje med positiv lutning och om $r = -1$ på en rät linje med negativ lutning. Om r ligger nära 0 så tyder det på att det inte finns något *linjärt* samband mellan variablerna.

Beroendemått



För standardmaterialet får vi $r = 0.95$ och för testmaterialet får vi $r = 0.11$.

Beroendemått

Om r ligger nära -1 eller 1 så tyder det på att det finns ett starkt linjärt samband mellan variablerna.

Men: samband är inte detsamma som orsakssamband!

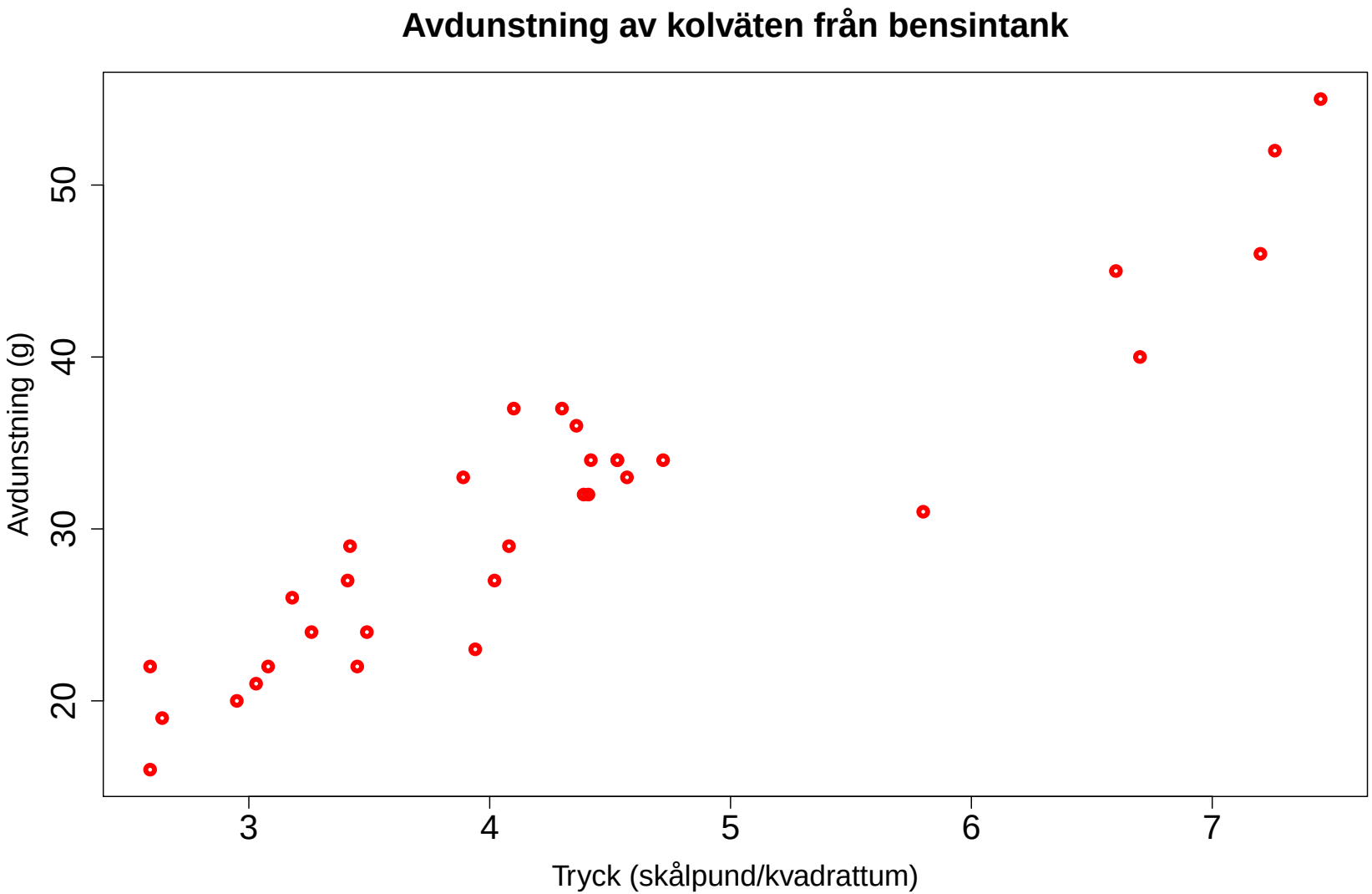
Exempel: vid en undersök av barns läs- och skrivförmåga upptäckte man att barn med stora fötter stavar bättre.

Exempel: glassförsäljning per månad och antal drunkningsolyckor per månad har hög korrelation.

Exempel: för perioden 1945-1957 så är korrelationen mellan antalet häckande storkar i Köpenhamn och antalet barn som föddes i staden hög.

Exempel: tryck i bensintank

När man tankar en bil med bensin så avdunstar kolväten. En biltillverkare undersökte mängden kolväten som avdunstade vid olika tryck i bensintanken:



Linjära regressionsmodeller

En *regressionsmodell* är en matematisk modell som beskriver ett samband mellan en *responsvariabel* y och *förklarande variabler* $x_1, \ldots, x_k$:

$$y = f(x_1, \ldots, x_k).$$

Om f är en linjär funktion så har vi en *linjär regressionsmodell*:

$$y = m + k_1 x_1 + k_2 x_2 + \ldots + k_k x_k.$$

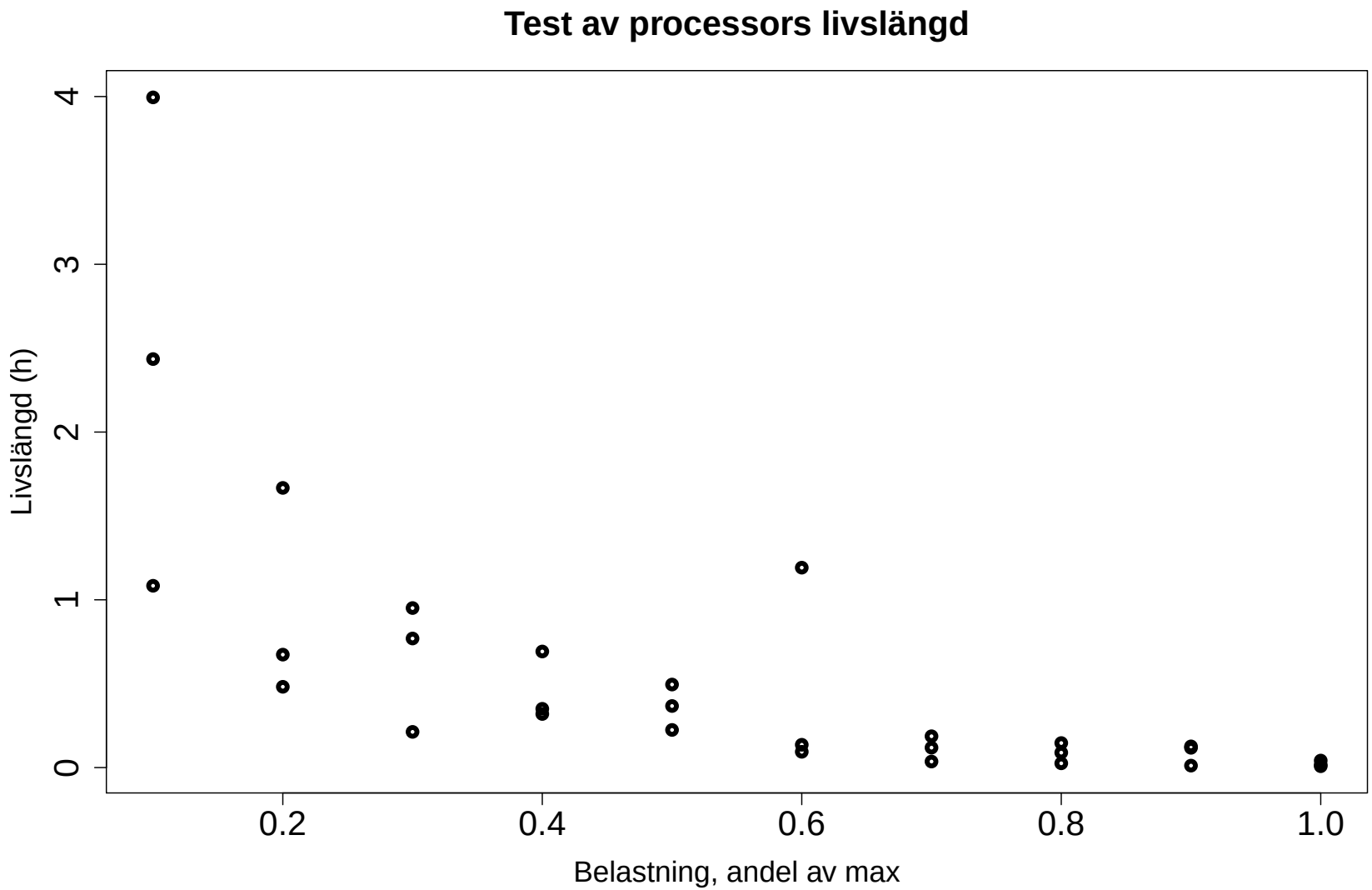
Vi ska här studera *enkla linjär regressionsmodeller*, där det bara finns en förklarande variabel x:

$$y = kx + m + \epsilon$$

där ϵ är en slumpavvikelse från den linjära trenden.

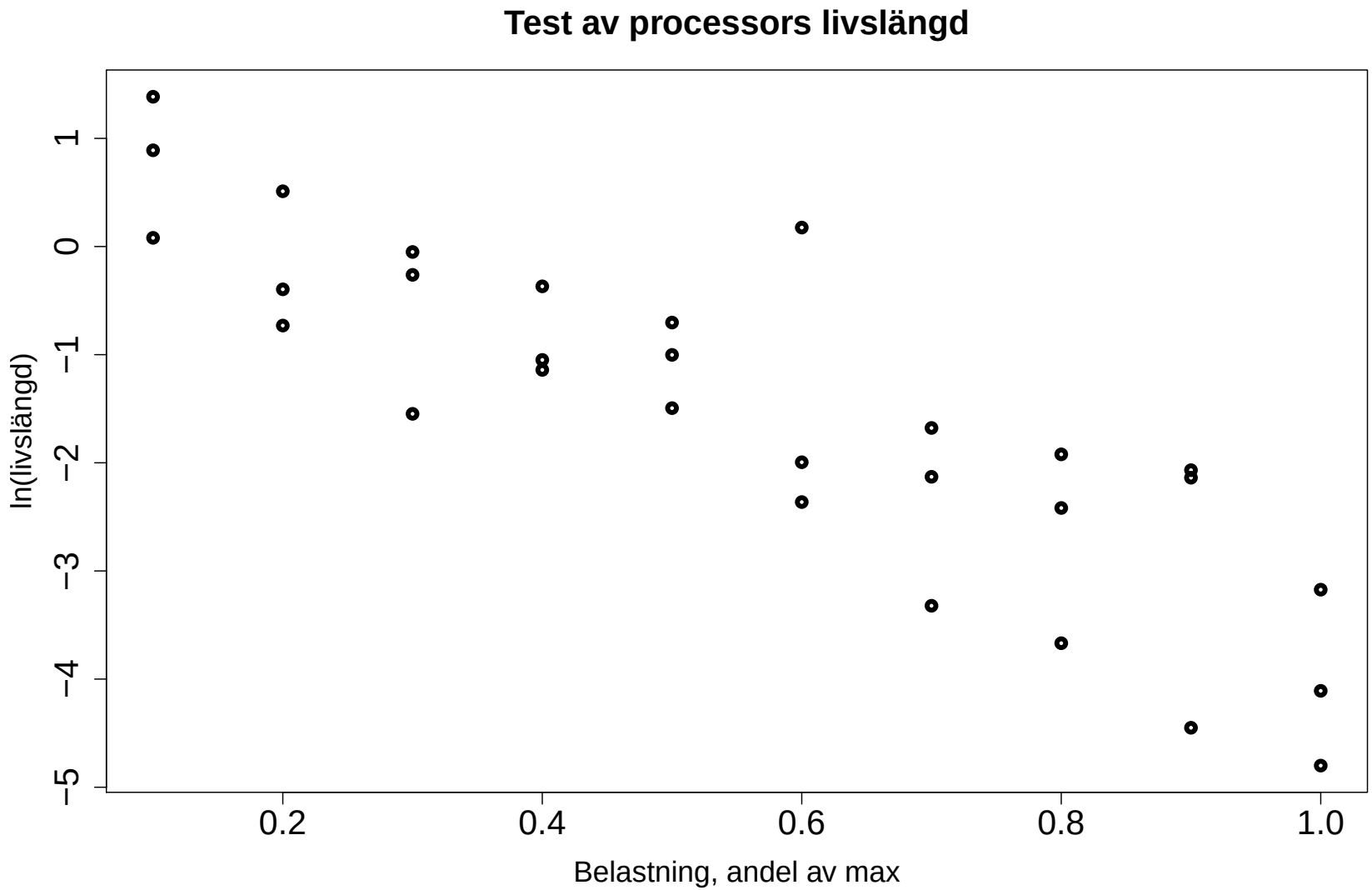
Varför linjära funktioner?

En datortillverkare undersökte livslängden hos en experimentell processor vid olika belastningar.



Varför linjära funktioner?

Datortillverkaren fann att ett samband av typen $y = ae^{bx}$ rådde.



Användningsområden

$$y = kx + m$$

- Skattningar av m och k ger en linje som kan använda för att uppskatta värden på y för nya värden på x.
- Sådana uppskattningar kallas *prediktioner*.

Linjära regressionsmodeller och prediktioner används exempelvis för

- Att undersöka samband mellan variabler
- Prognosmakande
- Kalibrering
- Reliabilitetsanalys
- Processoptimering

Stokastisk modell

Vi har n observationer av talparen (x_i, y_i) och vill undersöka om det finns ett beroende av typen

$$y = kx + m.$$

Antag att värdena $x_1, \ldots, x_n$ är fixa och bestämda på förhand medan $y_1, \ldots, y_n$ varierar slumpmässigt.

För att göra sambandet stokastiskt så antar vi följande modell:

$$Y_i = kx_i + m + \epsilon_i$$

där $\epsilon_i \sim N(0, \sigma^2)$ och olika ϵ_i är oberoende.

Alternativt kan detta skriva som att

$$Y_i \sim N(kx_i + m, \sigma^2).$$

Variationen mellan olika y_i antas alltså dels bero på deterministiska faktorer (olika x_i-värden) och dels på slumpmässiga faktorer (ϵ_i: mätfel, slumpavvikelser...).

Metod för linjeanpassning

Metoden som används för att anpassa den räta linjen till det givna datamaterialet kallas *minsta kvadratmetoden*.

Målet är att minimera följande uttryck med avseende på m och k:

$$\sum_{i=1}^{n}(y_i - (kx_i + m))^2$$

- Se tavlan!

Uttrycket kan minimeras genom att man deriverar det med avseende på m respektive k och undersöker för vilka värden på parametrarna som uttrycket är 0.

Metod för linjeanpassning

Låt

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

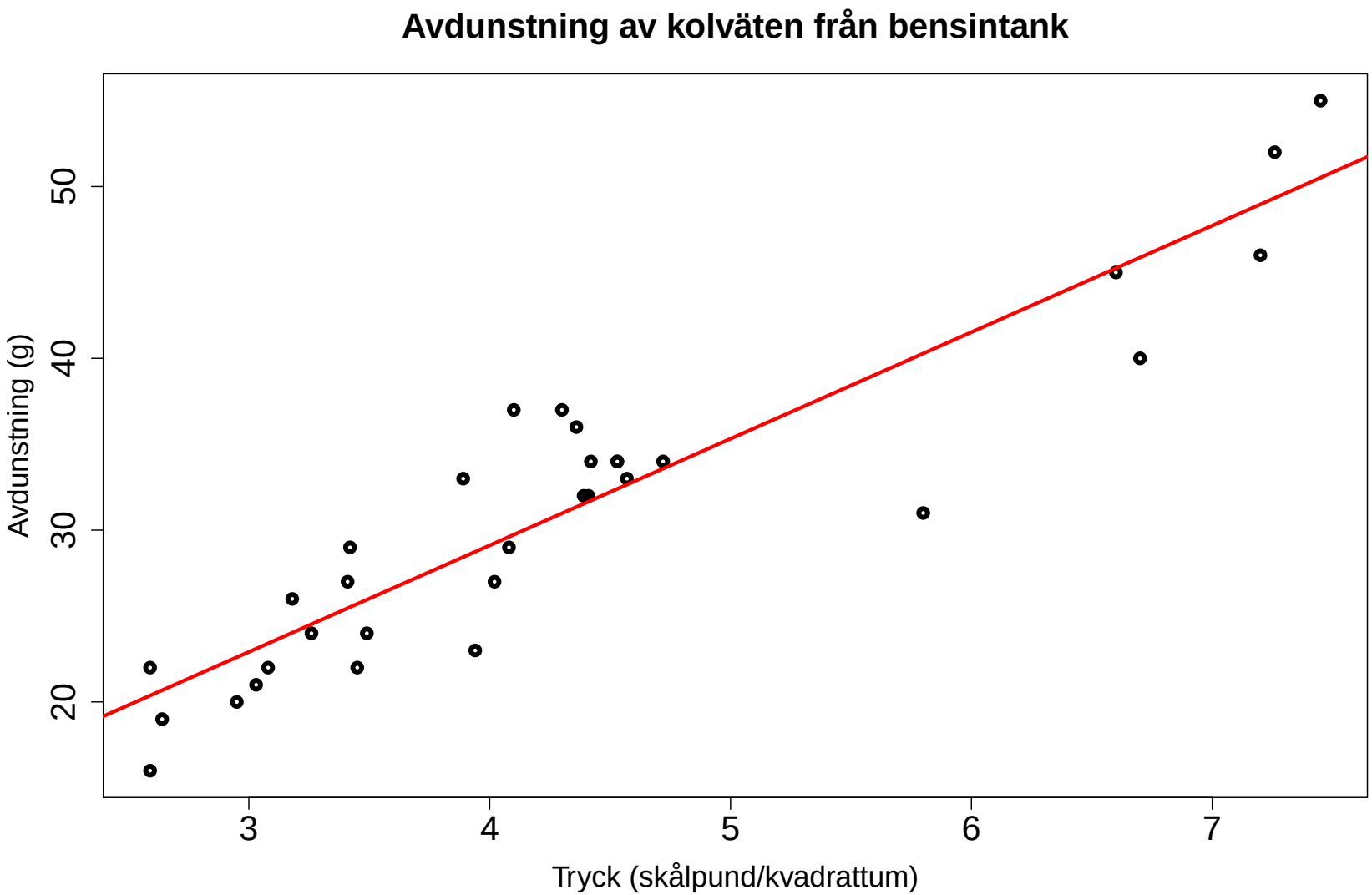
Derivering av uttrycket på föregående sida ger skattningarna

$$\hat{k} = \frac{S_{xy}}{S_{xx}} \qquad \text{och} \qquad \hat{m} = \bar{y} - \hat{k}\bar{x}.$$

Vidare kan variansen σ^2 skattas som

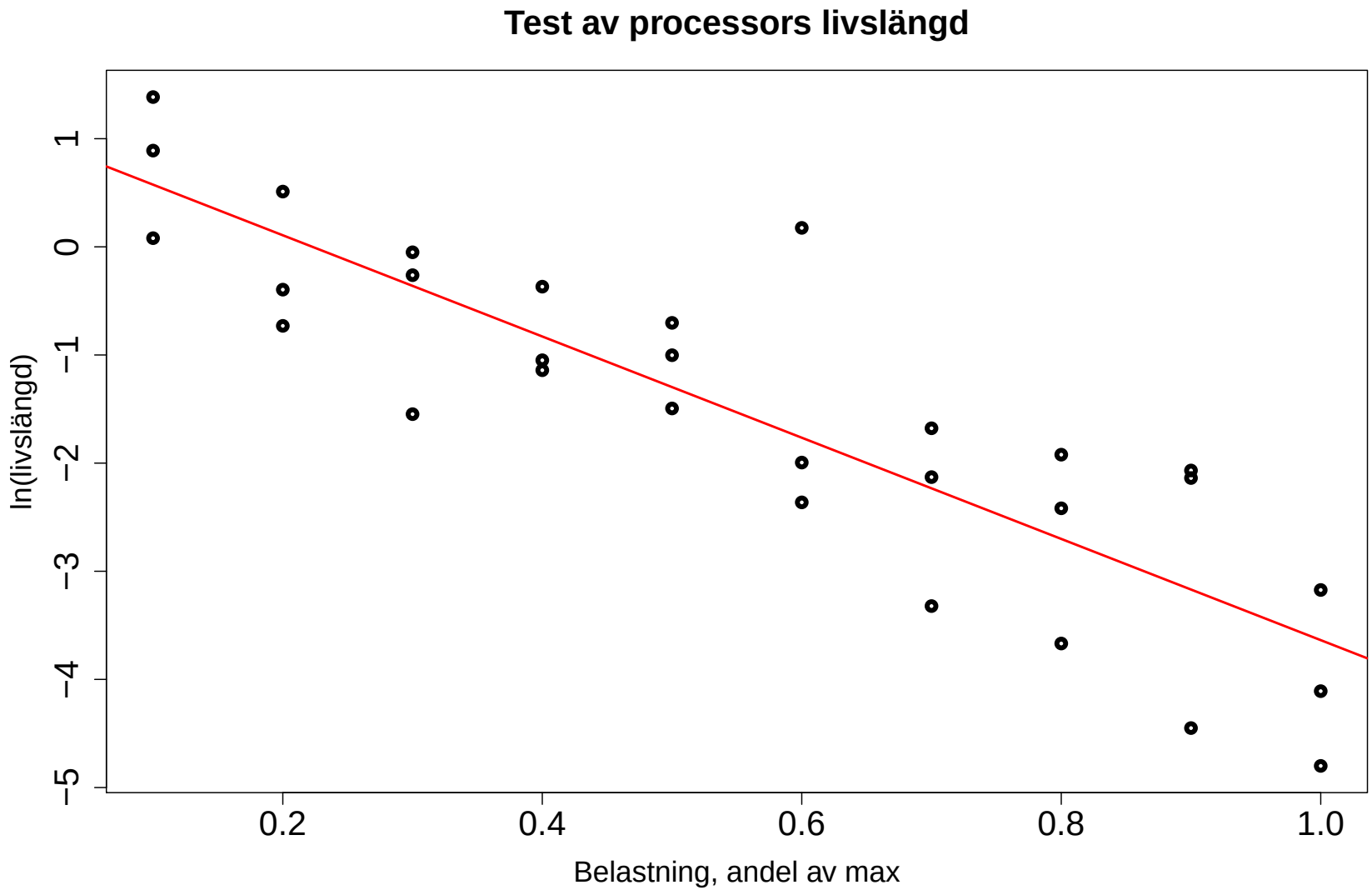
$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right).$$

Exempel: tryck i bensintank



Exempel: processorbelastning

För processorbelastningsdata fann man $\hat{m} = 1.0$ och $\hat{k} = -4.7$.



Hur starkt är sambandet?

Vi har tidigare beräknat *korrelationskoefficienten*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

som beskriver i vilken utsträckning det finns ett linjärt samband mellan variablerna x och y.

Inom regressionsanalys brukar *förklaringsgraden* $R^2 = r^2$ användas som ett mått på hur väl modellen $y = kx + m$ beskriver observerade data.

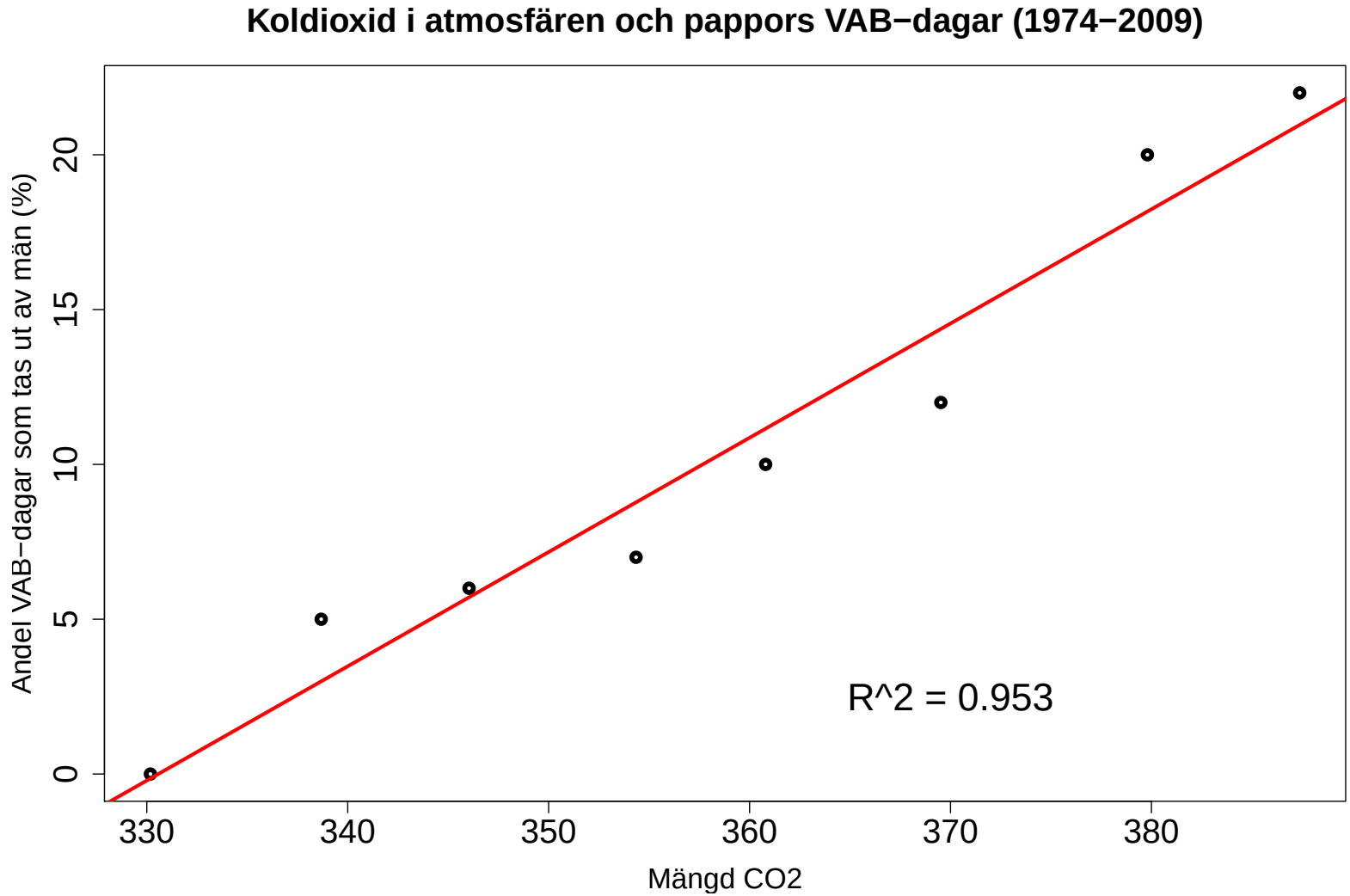
Det gäller att $0 \leq R^2 \leq 1$, där höga värden på R^2 tyder på att anpassningen är bra och $R^2 = 1$ betyder att alla värden ligger precis på linjen.

Att R^2 är nära 0 tyder antingen på att $k = 0$ (d.v.s. det finns inget samband) eller att sambandet inte är linjärt.

- Se tavlan!

Varnande exempel: koldioxid och Vård Av (sjukt) Barn

Sedan 1974 har mängden koldioxid i atmosfären ökat. Det har även andelen VAB-dagar som tas ut av män.



Får den ökade mängden koldioxid män att ta ut fler VAB-dagar...?

Konfidensintervall för k

Under antagandet att $\epsilon_i \sim N(0, \sigma^2)$ så kan man visa att

$$\hat{m} \sim N\left(m, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right)$$

och

$$\hat{k} \sim N\left(k, \frac{\sigma^2}{S_{xx}}\right).$$

Ur det senare uttrycket kan man härleda ett konfidensintervall för k.

Då σ^2 skattas med $s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$ fås konfidensintervallet

$$I_k = (\hat{k} \pm t_{\alpha/2}(n-2) \cdot s/\sqrt{S_{xx}})$$

som har konfidensgrad $1 - \alpha$.

- Se exempel på tavlan!

Hur ska konfidensintervallet tolkas?

Prediktion och prediktionsintervall

Givet skattningarna $\hat{m}$ och $\hat{k}$ så får vi ett predikterat värde på y_0 givet x_0:

$$y_0^{(pred)} = \hat{m} + \hat{k}x_0.$$

Dock finns en viss osäkerhet i prediktionen eftersom vi inte har de sanna värdena på parametrarna utan bara skattningar. Vi kan ta hänsyn till osäkerheten i skattningarna och skapa ett konfidensintervall för $\mathrm{E}[Y_0]$:

$$I_{\mathrm{E}[Y_0]} = (\hat{m} + \hat{k}x_0 \pm t_{\alpha/2}(n-2)s\sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}).$$

Men... Precis som för andra y-värden så kommer y_0 troligen att avvika lite från linjen på grund av slumpavvikelser.

Man kan därför slutligen beräkna ett *prediktionsintervall*, som är ett konfidensintervall för värdet på y_0 där både skattningarnas osäkerhet och osäkerheten orsakad av slumpavvikelsen för y_0 har inkluderats:

$$I_{y_0} = (\hat{m} + \hat{k}x_0 \pm t_{\alpha/2}(n-2)s\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}).$$

Prediktion och prediktionsintervall

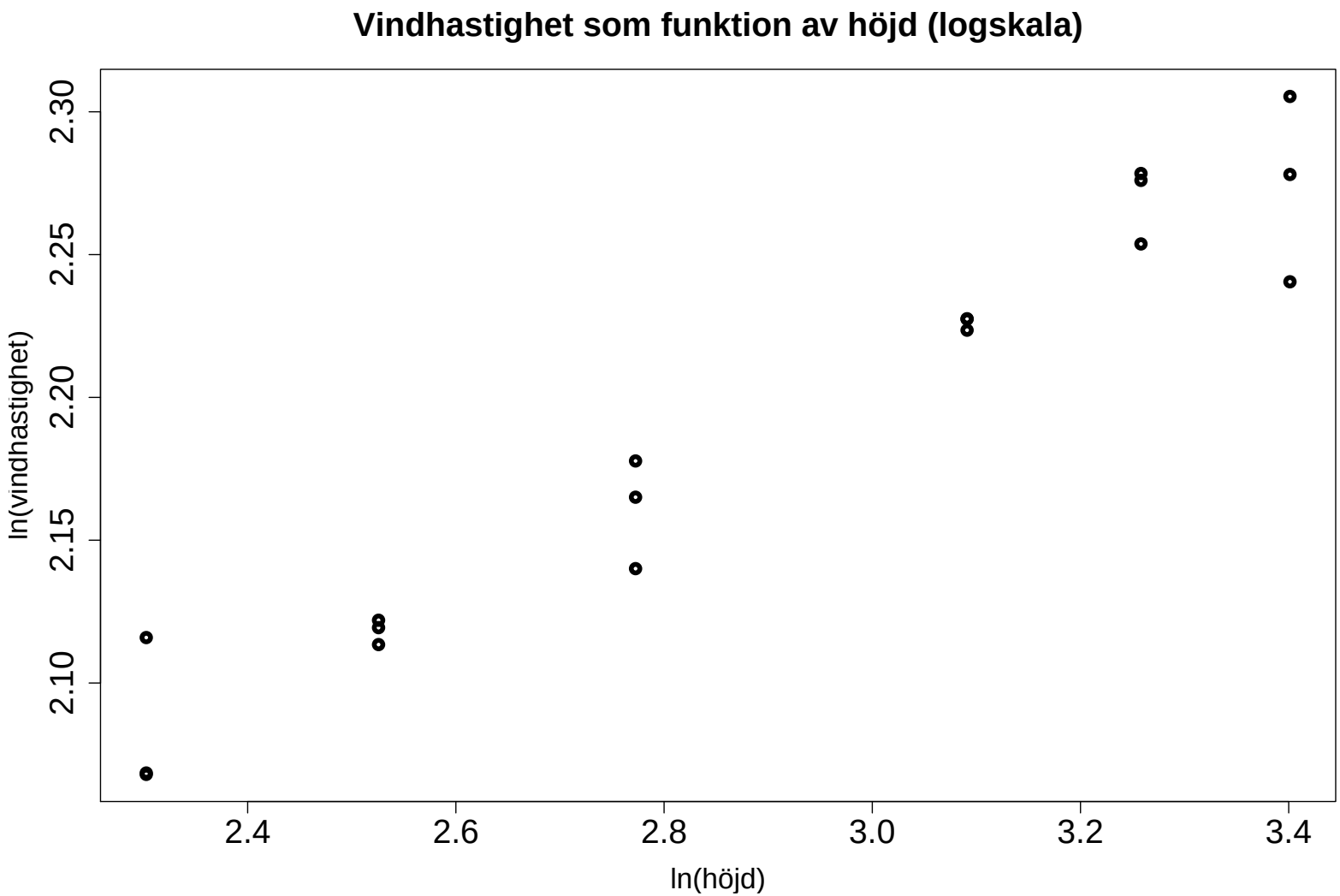
Prediktion: $y_0^{(pred)} = \hat{m} + \hat{k}x_0$, skattning av vilket y-värde som x_0 borde ge.

Konfidensintervall för $\mathrm{E}[Y_0]$**:** när man tar hänsyn till osäkerheten i skattningarna $\hat{m}$ och $\hat{k}$. Ett intervall med troliga värden på kurvan.

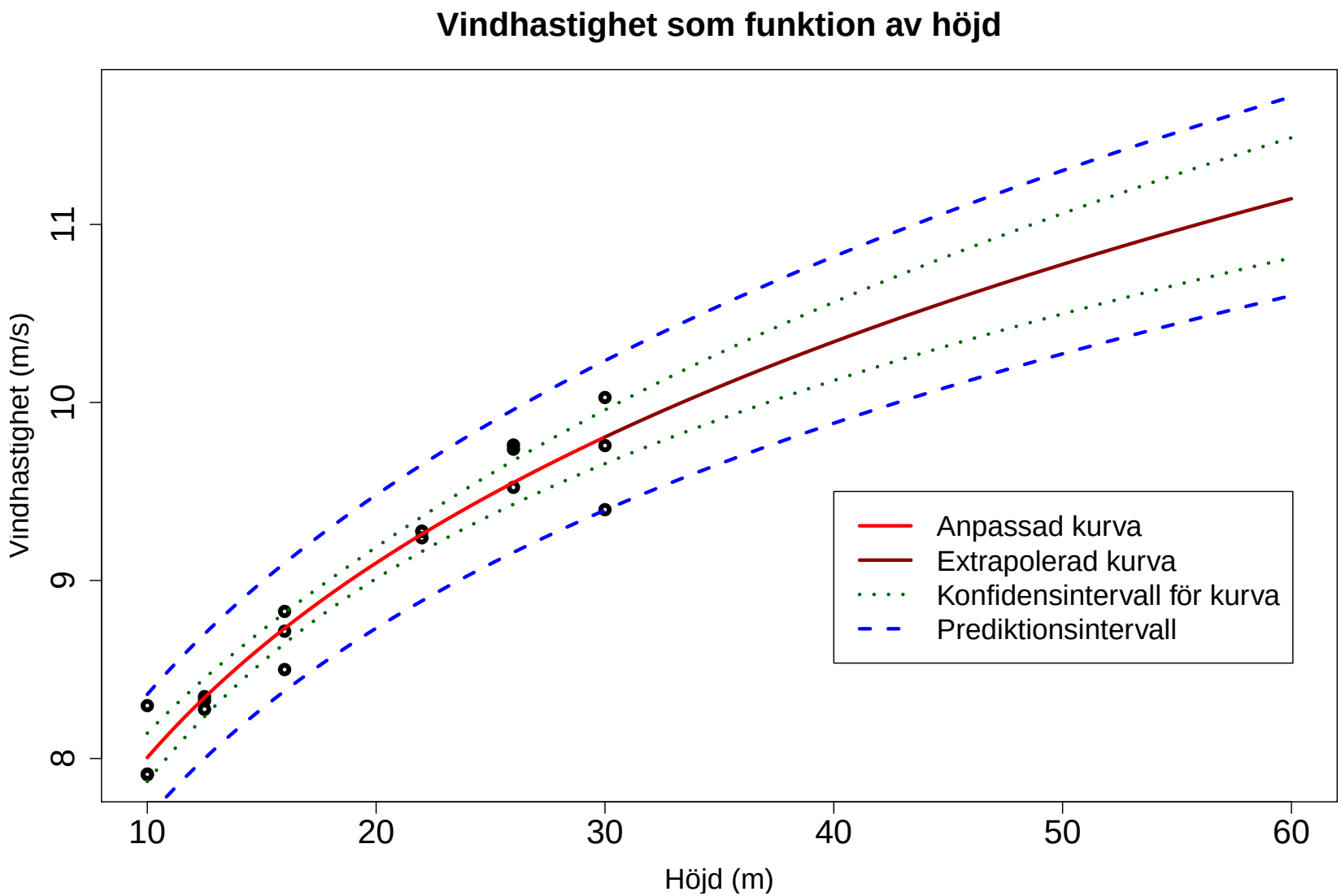
Prediktionsintervall: när man också tar hänsyn till att y_0 kommer att avvika lite från linjen på grund av slumpvariation. Ett intervall med troliga värden på y_0.

Prediktion: vindhastigheter

På en plats vill man ta reda på den genomsnittliga vindhastigheten vid 60 m höjd. Detta är svårt att mäta direkt, men man kan mäta vid 10-30 m höjd.



Prediktion: vindhastigheter



Extrapolering

Även om den linjära modellen ger en bra beskrivning av ett samband i det område som undersökts så behöver det inte vara så att sambandet är linjärt även utanför detta område.

VARNING! Det kan vara farligt att använda modellen utanför det område där den anpassats!

- Se exempel på tavlan!

Sammanfattning

- Linjära regressionsmodeller
- Stokastisk modell
- Linjeanpassning och skattningar
- Konfidensintervall
- Prediktion och extrapolering