Statistisk översiktskurs - Föreläsning 8

Anders Fredriksson

Statistiska Institutionen
Stockholms Universitet

7 april 2025

Föreläsning 8 - innehåll

- Fortsättning enkel linjär regression, teori
- Tolkning av regressionsresultat
  - Regressionskoefficienter
  - R2
  - Residualer
- Prediktion
- Inferens och ytterligare tolkning av regressionsresultat
  - t-värde och p-värde
  - Konfidensintervall

Enkel linjär regression - populationen

- Vi vill förstå hur ett samband ser ut i populationen
- **Ex:** Samband mellan lägenhetsstorlek (X) och hyresnivå (Y) i Uppsala
- **Population** - alla lägenheter i Uppsala
- Sambandet i populationen modelleras som
  $Y = \beta_0 + \beta_1 X + \varepsilon$
- "Hyran är lika med en konstant (β_0 - "beta0") plus en lutnings-koefficient (β_0 - "beta1") gånger ytan plus en felterm (ε - epsilon)."
- Räta linjens ekvation, plus en felterm som antas vara noll i medel
- Grekiska bokstäver används typiskt för populationssamband

- Problem: Vi har typiskt inte data på hela populationen och kommer inte att kunna få reda på **populationssambandet**

Enkel linjär regression - vårt urval/stickprov

- **Ex: Urval** - lägenheter i Uppsala (förhoppningsvis valda slumpvis)
- Vi **skattar** förhållandet i populationen genom att, **för vårt urval**, ta fram b_0 och b_1 i följande relation, med minsta kvadratmetoden,

  $Y = b_0 + b_1 X + e$
- Minsta kvadratmetoden - minimera e^2 (F7, s. 15-16)
- b_0 är en skattning av β_0
- b_1 är en skattning av β_1
- e är felet i vår skattning (residualen)

- $\hat{y} = b_0 + b_1 x$ är skattningen av y för ett visst x-värde (y-hatt)

Regressionsresultat från vårt exempel i F7

```
Call:
lm(formula = hyra ~ yta, data = Uppsaladata)

Residuals:
    Min      1Q  Median      3Q     Max
-693.28 -450.36  -70.95  364.44 1092.24

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  720.923    370.244   1.947   0.0665 .
yta           60.533      5.713  10.595 2.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 525.5 on 19 degrees of freedom
Multiple R-squared:  0.8553,    Adjusted R-squared:  0.8476
F-statistic: 112.3 on 1 and 19 DF,  p-value: 2.057e-09
```

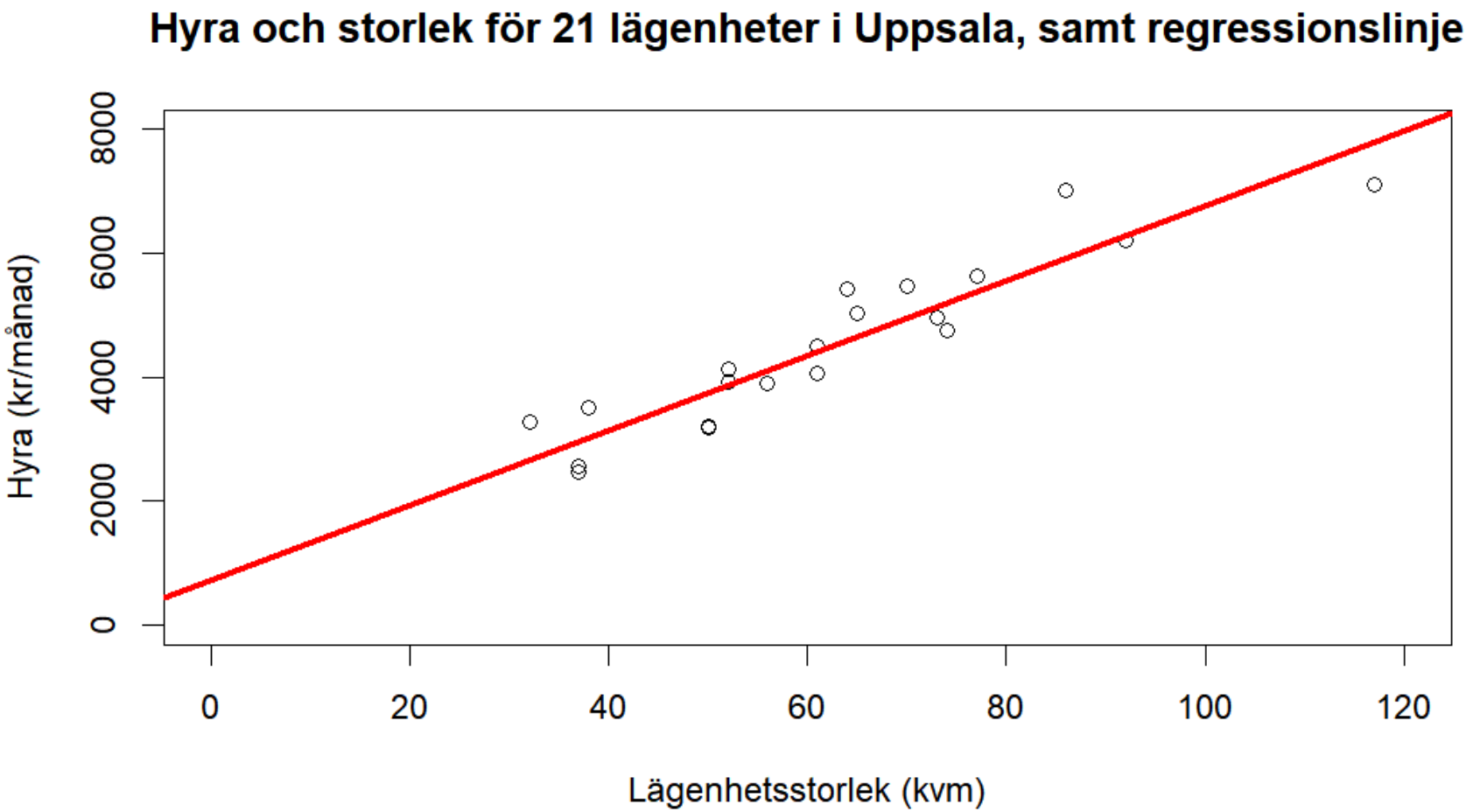
Regressionskoefficienter

t-värde

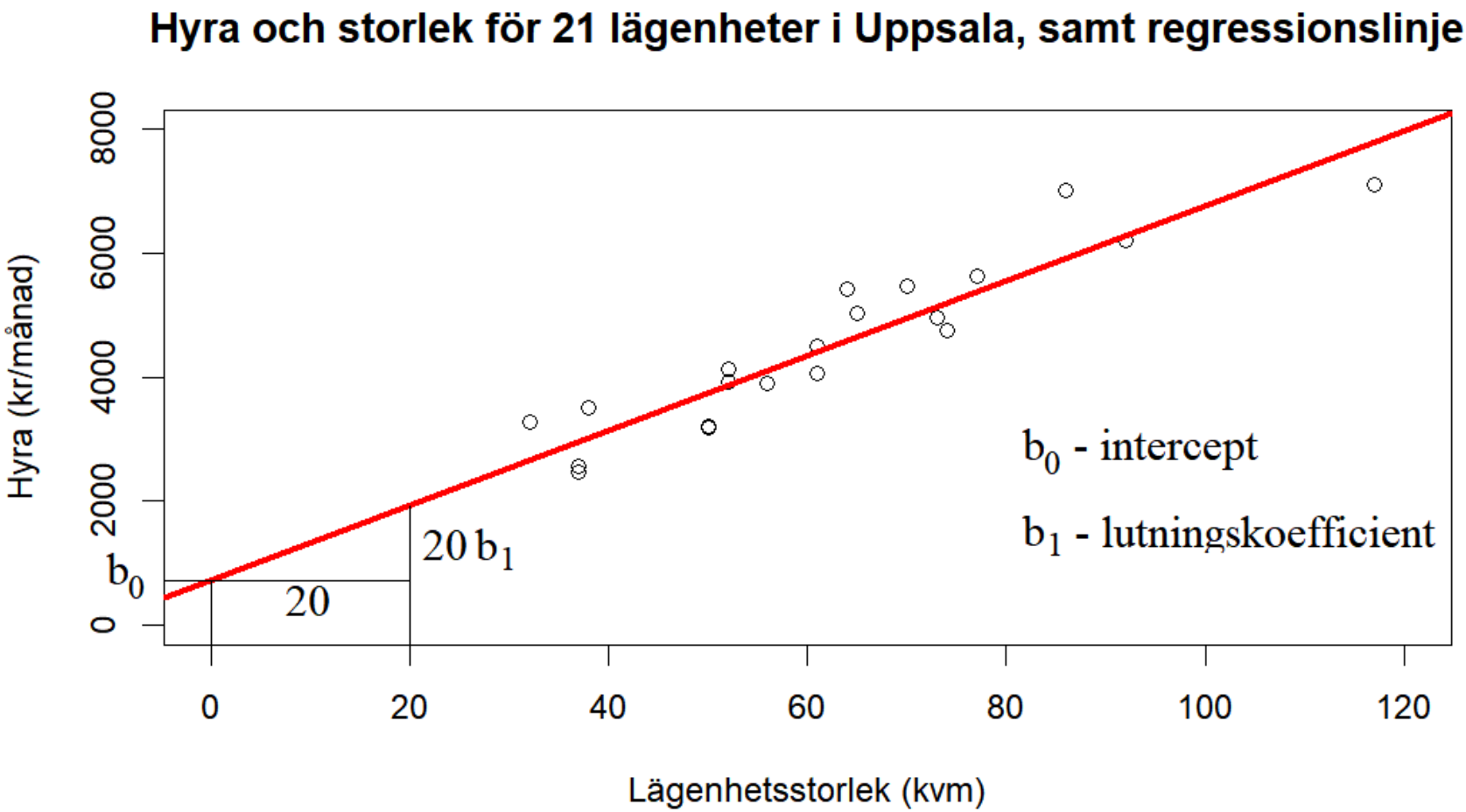
p-värde

R2

Skattad regressionslinje



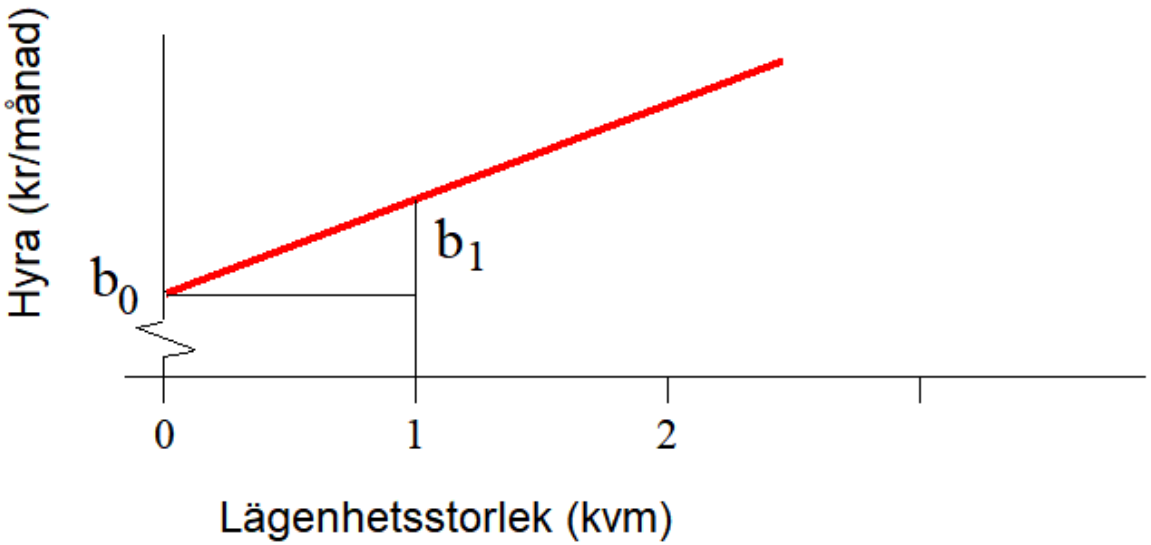
Intercept och lutningskoefficient



Förstora upp x-axeln - tolkning

b_0 - intercept - en lägenhet på 0 kvm skattas ha en hyra b_0

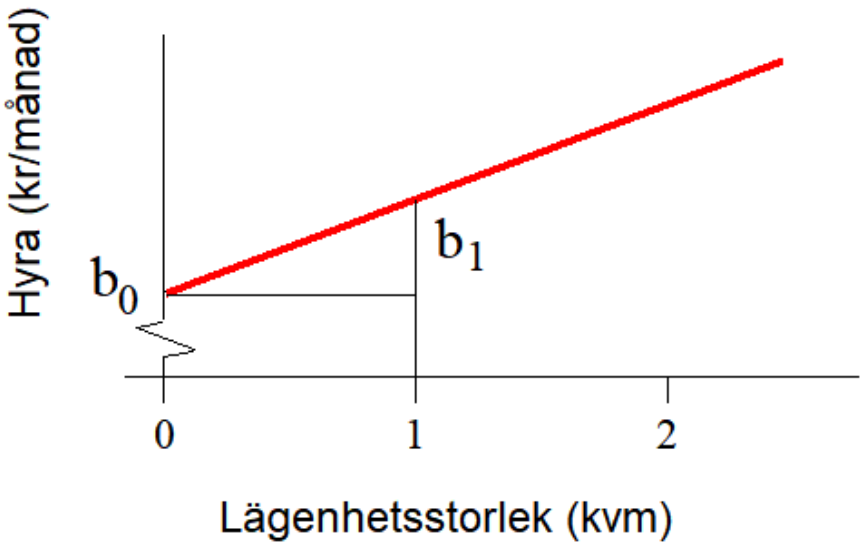
b_1 - lutningskoefficient - en enhets ökning i storlek skattas vara associerad med b_1 enheter högre hyra



Tolkning - sätt in värden och enheter

b_0 - intercept - en lägenhet på 0 kvm skattas ha en hyra a 721 kronor/månad

b_1 - lutningskoefficient - en enhets ökning i storlek, dvs. en ökning med 1 kvm, skattas vara associerad med 60,5 kr/månad högre hyra



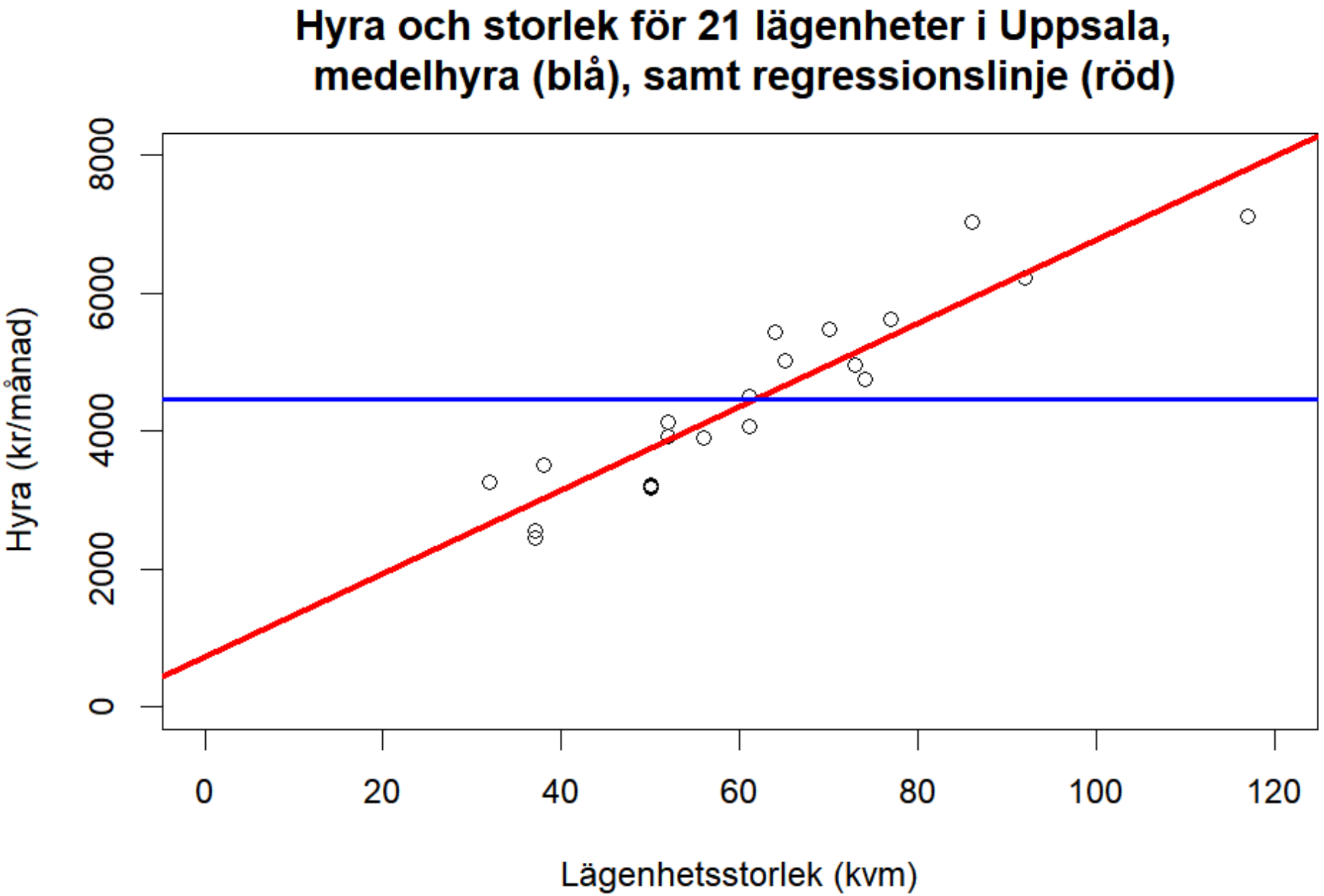
Våra skattningar:

```
Coefficients:
             Estimate
(Intercept)  720.923
yta           60.533
```

Enkel linjär regression – R2

- **R2 (eller R^2, R-kvadrat)** (R-squared), används som ett mått på hur mycket en regressionsmodell förklarar av variationen i responsvariabeln.
- Om vi inte hade någon modell och inga förklaringsvariabler och någon bad oss att skatta y skulle vår bästa skattning vara $\bar{y}$ (medelvärdet)
- Nu har vi istället en förklaringsvariabel och en modell som förklarar en del av variationen i y mha x
- R2 är ett mått på hur stor del av variationen i y som förklaras
- R2 kommer alltid att vara ett värde mellan 0 och 1
- Bilden: Vår modell (röd linje) förklarar en stor del av variationen i y

Skatta y med medelvärde vs. skatta y med modell



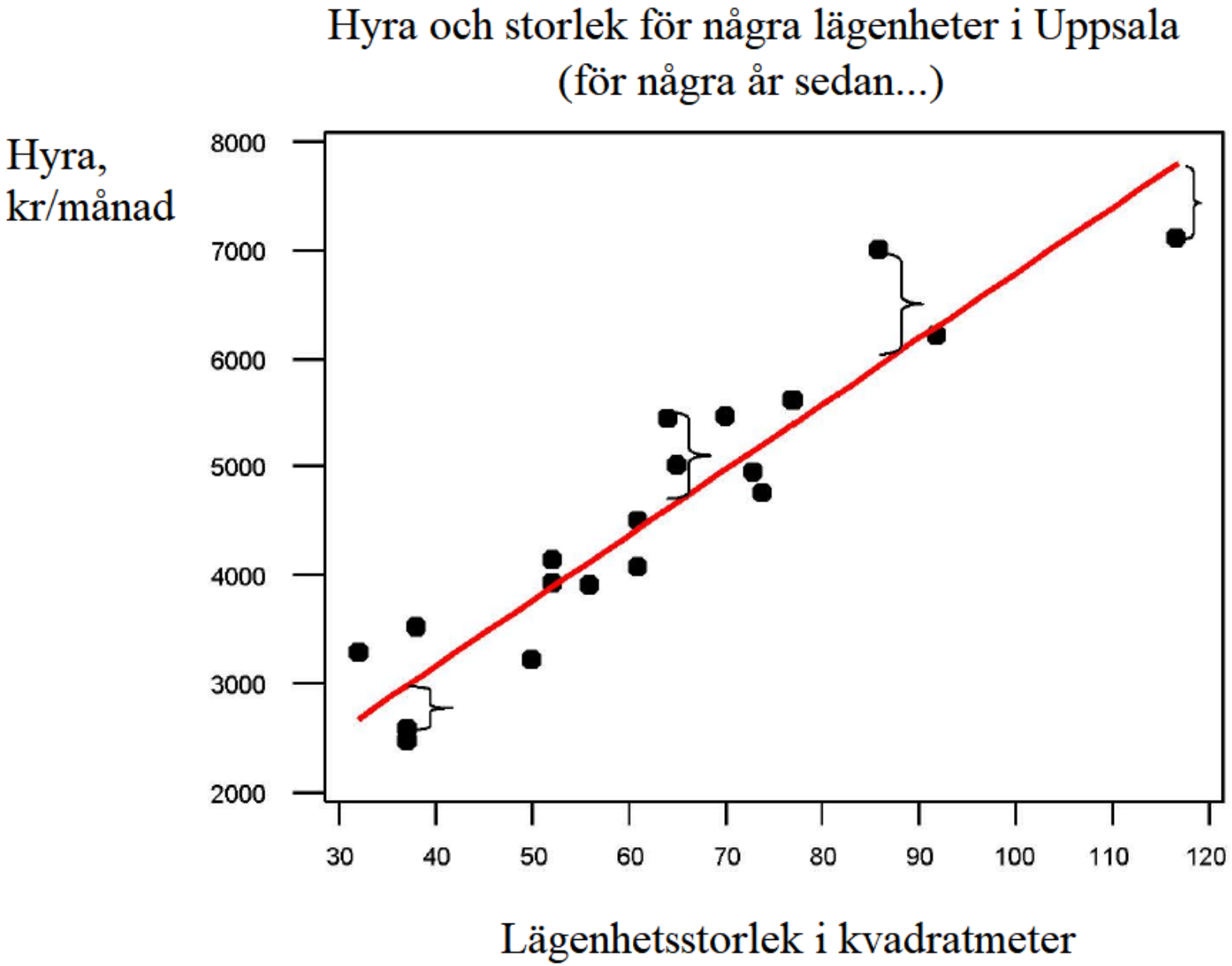
R2:s betydelse

- För enkel linjär regression gäller att $R^2 = r^2$, där r är korrelationskoefficienten mellan X och Y. (F7, s. 12))
- Det går inte att säga generellt vad som är ett "bra" värde på R^2
- Inom naturvetenskaper kan R^2 ofta vara nära 1
- Inom samhällsvetenskaper är R^2 ofta under 0.5, eftersom många faktorer påverkar och modeller sällan fångar alla faktorer

Residualer (residuals)

- Residualen för varje observation är y-värdet minus det skattade y-värdet, $e = y - \hat{y}$
- Residualen kan tolkas som avståndet i y-led från regressionslinjen (negativ om observationen ligger under linjen)
- Residualerna är i medel noll
- Vi analyserar residualerna bland annat för att avgöra om vi har skattat en lämplig modell
- Om våra data exv. inte har ett linjärt mönster, kommer vi se detta i residualanalysen, i residualplottar
- Se boken kap. 7, exv. dataset 2, sid. 111
- Mer residualanalys i andra statistikkurser, viktiga antaganden, etc.

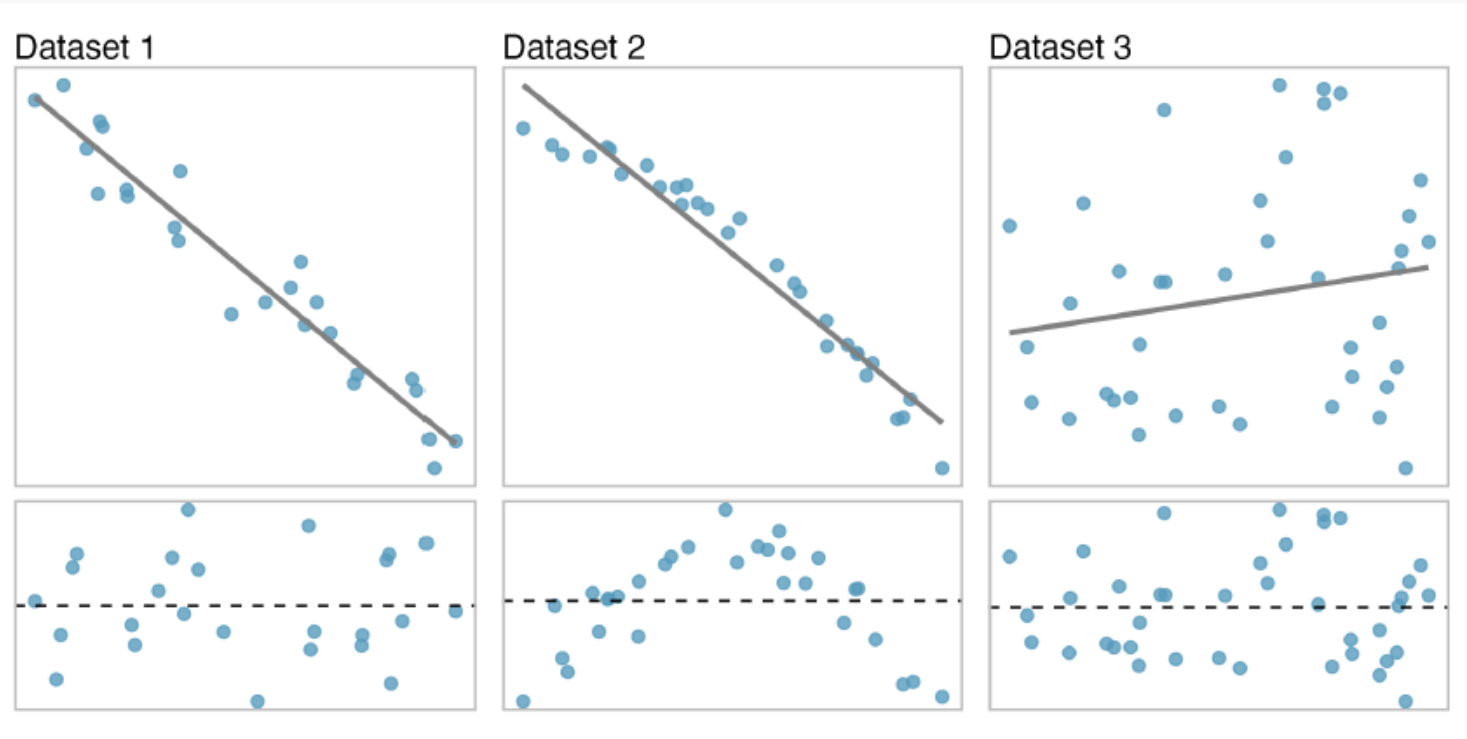
Residualer, exempel, Uppsala



Residualplottar, exempel, boken

EXAMPLE

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. The figure below shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns in the residuals?



Prediktion (lite mer om begreppet)

- Att **prediktera** (predict) är att göra en uppskattning av ett värde när vi inte kan göra en direkt observation.
- Substantivformen av ordet är prediktion. Exempel: "Syftet med vår modell är att göra en prediktion av bensinförbrukningen."
- Vi kan också använda orden **estimera eller skatta** (estimate) med ungefär samma betydelse som prediktera. Exempel: "Vi estimerar/skattar bensinförbrukningen till 1.7 liter per mil."
- Substantivformerna är estimat och skattning. Exempel: "Vårt estimat/Vår skattning är att bilen förbrukar 1.7 liter per mil."

Prediktion, exempel

- Vi har skattat sambandet mellan lägenhetsstorlek och hyresnivå i Uppsala
- För en viss lägenhetsstorlek x (i kvm) kan vi skatta hyresnivån (i kr/månad) med följande modell:
- $\hat{y} = b_0 + b_1 x = 720.923 + 60.533x$

- $\hat{y}(x = 50) = b_0 + b_1 \times 50 = 720.923 + 60.533 \times 50 = 3747,573$
- En lägenhet på 50 kvm skattas ha en hyra på 3748 kr/månad

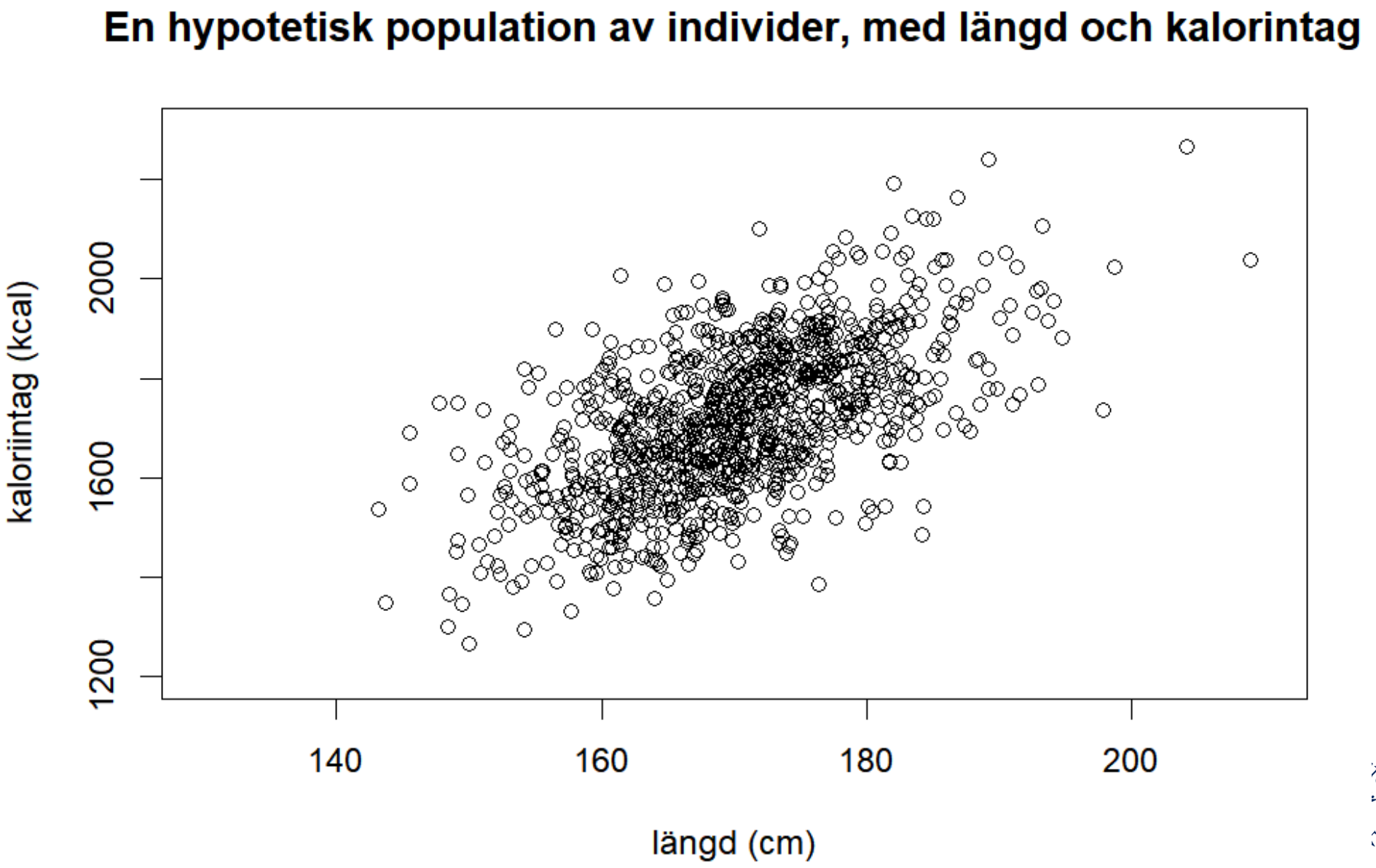
- $\hat{y}(x = 100) = b_0 + b_1 \times 100 = 720.923 + 60.533 \times 100 = 6774,223$
- En lägenhet på 100 kvm skattas ha en hyra på 6774 kr/månad
- Glöm inte enheter!

Inferens

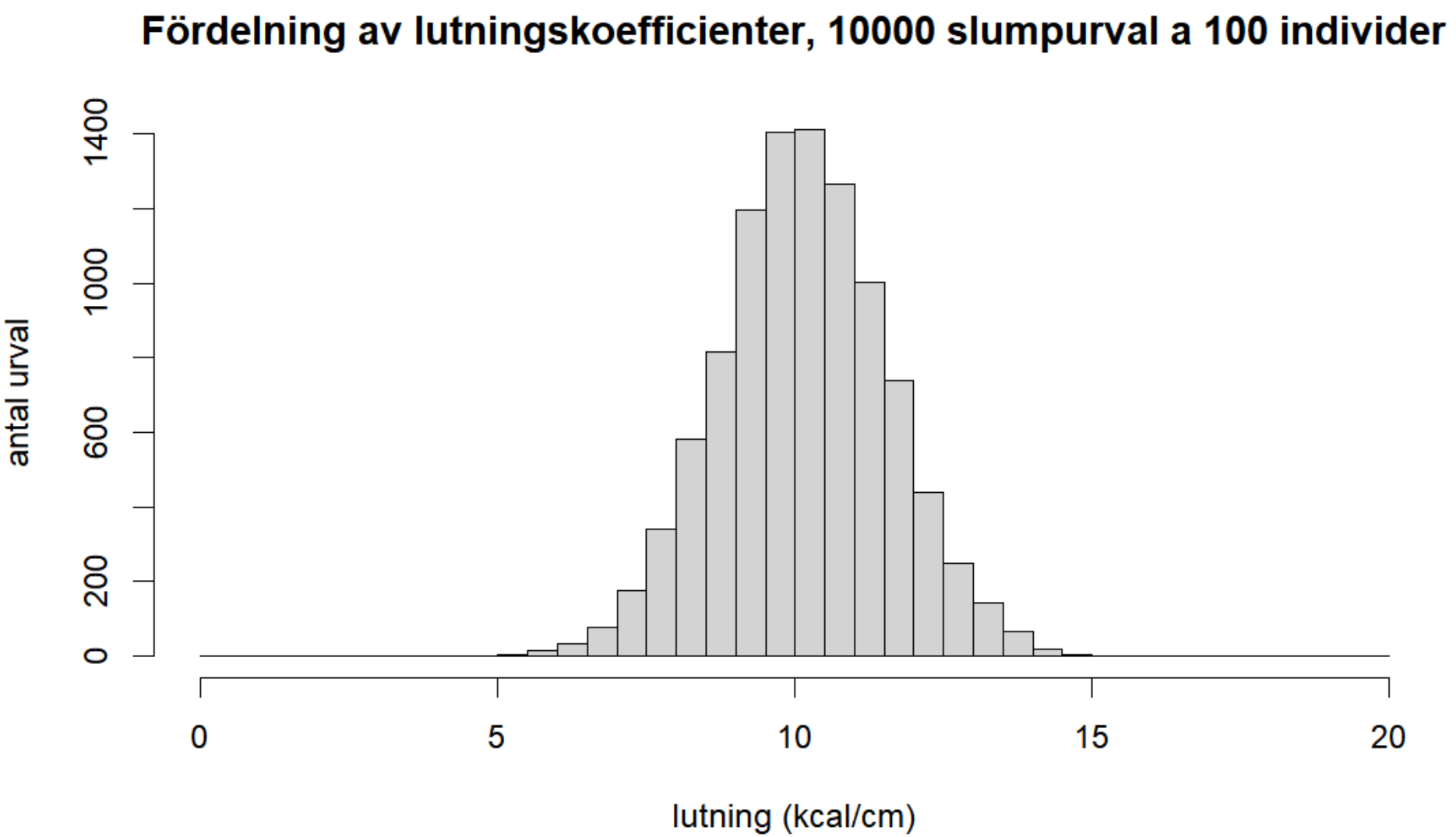
- På F4 och F7 såg vi exempel på hur lutningskoefficienten för sambandet mellan Y och X kan variera (olika urval från samma population)
- Vi kunde inte ta fram hela samplingfördelningen på lutningskoefficienter men vi drog 10000 slumpurval och fick fram en fördelning av lutningskoefficienter (se nästa bild)

- Mer generellt: Vilka slutsatser, avseende populationen, kan vi dra från ett visst urval, exempelvis från vår Uppsalastudie?
  - Är lutningskoefficienten skild från noll? (mao, finns ett samband?)
  - Hur stor är lutningskoefficienten? (ta fram ett konfidensintervall) **Ex:** Vilken "effekt" har lägenhetsstorlek på hyresnivå i Uppsala?
- Vi repeterar lite material från föreläsningarna 4 och 6

R-exempel från F4, F7



R-exempel från F4, F7



Inferens - hypotestest

- Vi är typiskt intresserade av om det finns ett samband eller inte, mellan två variabler (dvs. lutningskoefficienten):
  - $H_0 : \beta_1 = 0$ (Det finns inget samband)
  - $H_A : \beta_1 \neq 0$ (Det finns ett samband)
- Vi behöver ta fram en "Z-score" för vårt lutningsestimat, dvs ta fram vårt estimatat på "standardnormalform" (F4, s.26; F6, s.22)

Z-score vid hypotestest

Vid hypotesprövning beräknas Z-score för en punktskattning som

$$Z = \frac{\text{punktskattning} - \text{nollvärde}}{\text{SE}}$$

där **SE** (standard error) är motsvarigheten till standardavvikelsen för punktskattningen, och **nollvärdet** kommer från påståendet i nollhypotesen.

- Med våra hypoteser: Nollvärdet = 0, och punktskattning och standardfel (SE) fås från regressionsoutput!

Inferens - hypotestest

```
lm(formula = hyra ~ yta, data = Uppsaladata)

Residuals:
    Min      1Q  Median      3Q     Max
-693.28 -450.36  -70.95  364.44 1092.24

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  720.923    370.244   1.947   0.0665 .
yta           60.533      5.713  10.595 2.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 525.5 on 19 degrees of freedom
Multiple R-squared:  0.8553     Adjusted R-squared:  0.8476
F-statistic: 112.3 on 1 and 19 DF,  p-value: 2.057e-09
```

- Vi får $Z = \frac{60.533}{5.713} \approx 10.59566$ - vi har också detta värde i vår output

Inferens - hypotestest

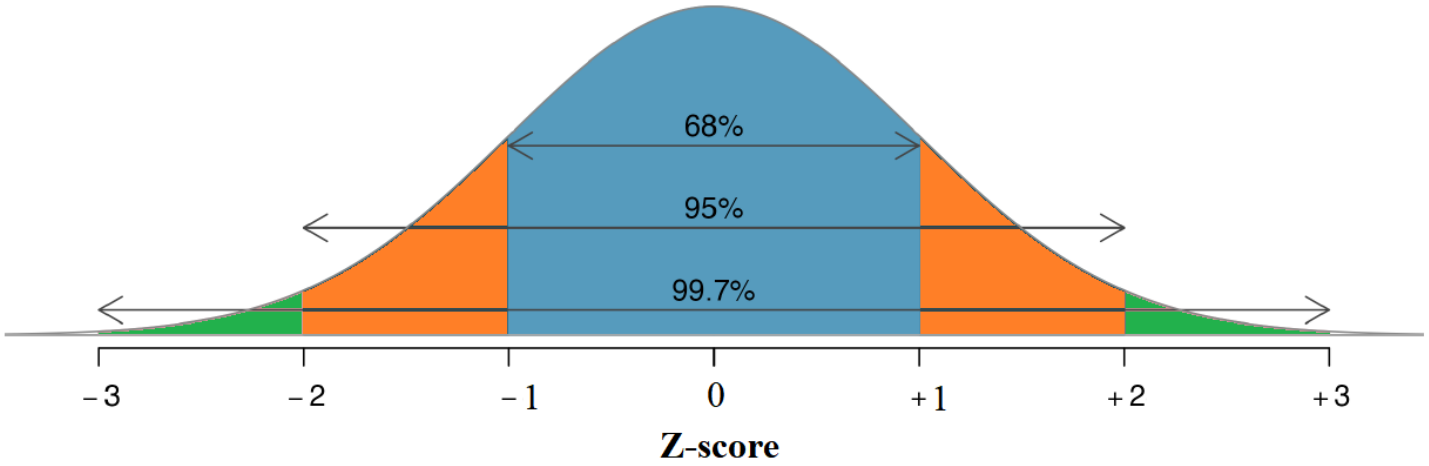
```
lm(formula = hyra ~ yta, data = Uppsaladata)

Residuals:
    Min      1Q  Median      3Q     Max
-693.28 -450.36  -70.95  364.44 1092.24

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  720.923    370.244   1.947   0.0665 .
yta           60.533      5.713  10.595 2.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 525.5 on 19 degrees of freedom
Multiple R-squared:  0.8553     Adjusted R-squared:  0.8476
F-statistic: 112.3 on 1 and 19 DF,  p-value: 2.057e-09
```

- Vi får $Z = \frac{60.533}{5.713} \approx 10.59566$ - vi har också detta värde i vår output
- Nästan...., vi har t-value, inte Z-value...

Inferens - t-fördelningen (kursivt / överkurs)

- Det faktum att vi inte vet standardavvikelsen i samplingfördelningen av regressionskoefficienter utan måste skatta denna - SE i regressionstabellen - gör att vi ska använda en t-fördelning istället för en normalfördelning
- För tillräckligt stora urvalsstorlekar är t-fördelningen mycket lik normalfördelningen, men lite "bredare i svansarna" - högre krav för att förkasta nollhypotesen
- Mest korrekt att skriva t-value istället för Z-value
- Vi får t-value$=\frac{60.533}{5.713} \approx 10.59566$
- Läs (kursivt) 19.2.3-19.2.5 i boken
- För vårt Uppsalaexempel använder vi normalfördelningen (även om n=21 är ett gränsfall)

Inferens - hypotestest
Bild från F4, F6 - normalfördelning, standardnormal form



- Om vi hade haft ett t-värde mellan -2 och 2 hade vi inte kunnat förkasta nollhypotesen, med 95% säkerhet
- Men vi har ett t-värde över 10 - vi kan med hög säkerhet förkasta nollhypotesen
- p-värdet i output (som är otroligt lågt) - sannolikheten att observera den lutningskoefficient vi gör, om nollhypotesen hade varit sann

Konfidensintervall för regressionskoefficient

- Liknar förfarandet på föreläsning 6, s. 9, se också boken 24.5:

Confidence intervals for coefficients.

Confidence intervals for model coefficients (e.g., the intercept or the slope) can be computed using the t-distribution:

$$b_i \pm t^\star_{df} \times SE_{b_i}$$

where $t^\star_{df}$ is the appropriate $t^\star$ cutoff corresponding to the confidence level with the model's degrees of freedom, $df = n - 2$.

- Vi kommer ta fram ett konfidensintervall på labb 4, för en lutningskoefficient
- De värden vi behöver är punktestimatet (b_1) och dess standardfel, båda finns i regressionstabellen
- För ett 95% konfidensintervall använder vi t-värdet 2 i formeln ovan

Denna version av dokumentet: 2025-04-07

Materialet i Statistisk översiktskurs har tagits fram av Ulf Högnäs och Anders Fredriksson, med inspiration och ibland direkt användande av material från andra kurser och personer, bland annat kurserna Statistik och dataanalys 1-3, med material av Michael Carlson, Ellinor Fackle Fornius, Jessica Franzén, Oskar Gustafsson, Oscar Oelrich, Mona Sfaxi, Karl Sigfrid, Mattias Villani, med flera.