F13 – Regression och problemlösning

Måns Thulin

Uppsala universitet

thulin@math.uu.se

Statistik för ingenjörer • 4/3 2013

Regression

Vi studerar hur en variabel y beror på en variabel x. Vår modell är

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

där ϵ är en slumpavvikelse från linjen $y = kx + m$. Vi antar att $\epsilon \sim N(0, \sigma^2)$.

Låt

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

Vi fick skattningarna

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

och

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

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

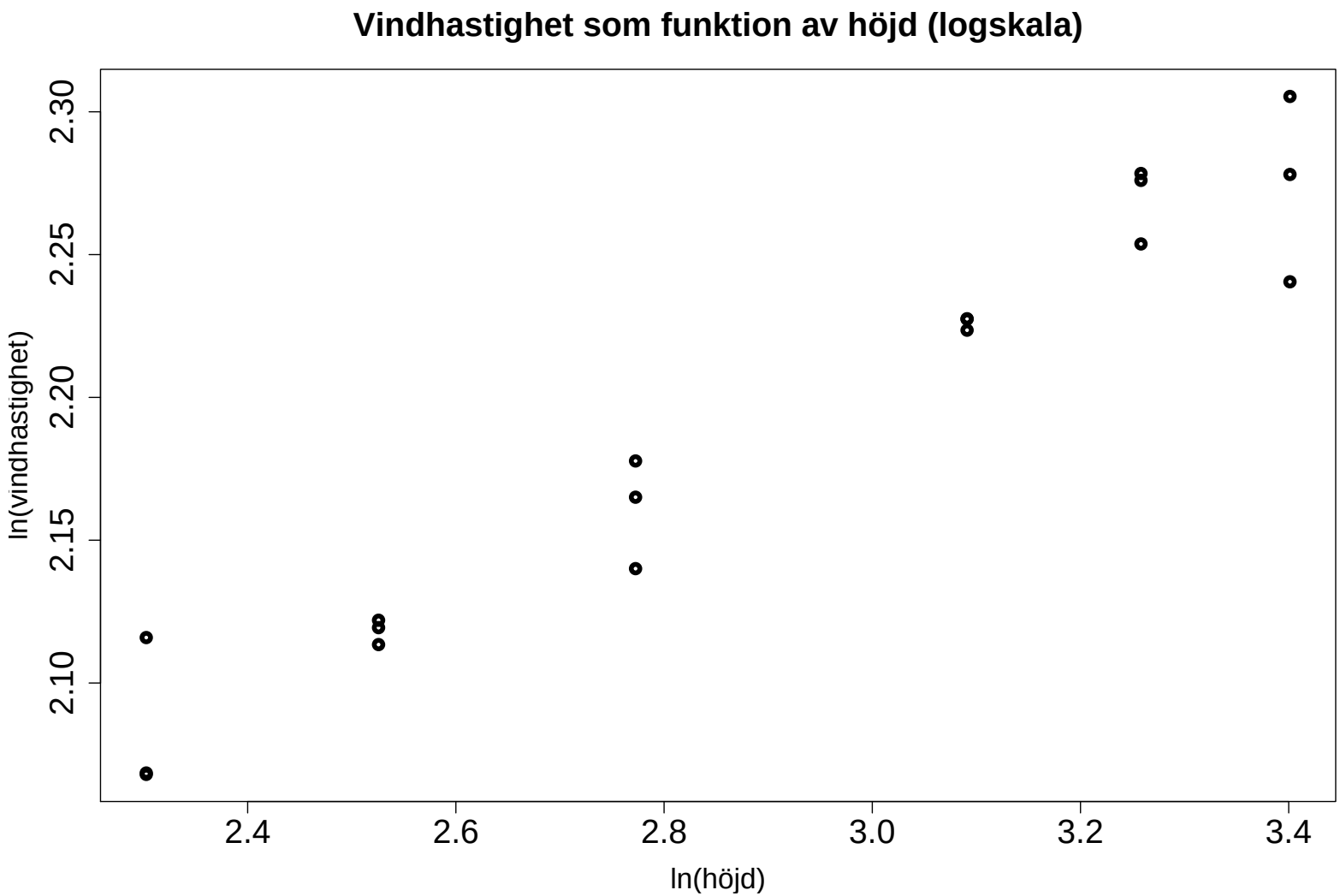
Prediktion: $y_0^{(pred)} = \hat{m} + \hat{k}x_0$, skattning av vilket y-värde som x_0 borde ge.

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.

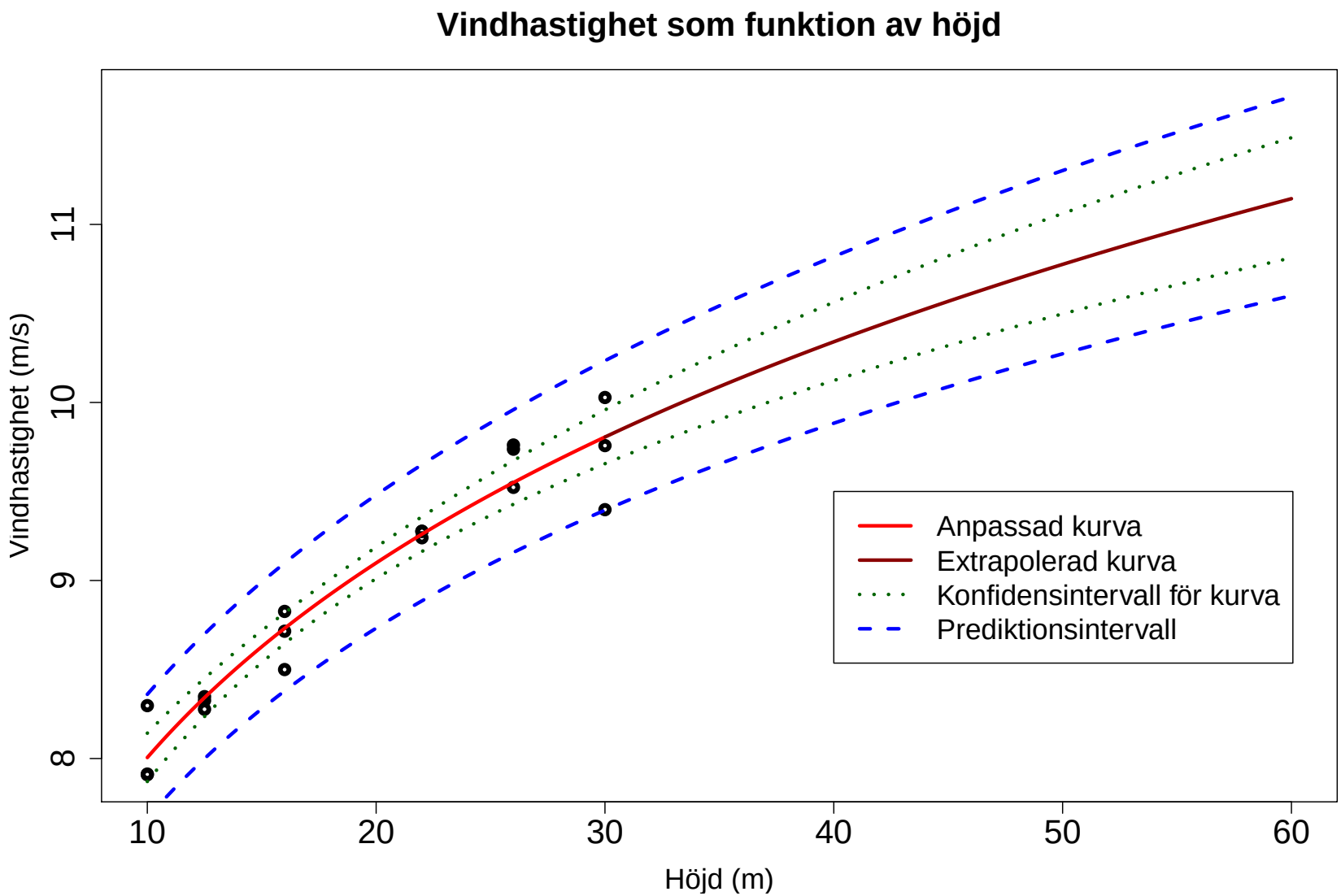
Prediktionsintervall: när man också tar hänsyn till att y_0 kommer att avvika lite från linjen på grund av slumpvariation. Ett intervall med troliga värden på y_0.

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!

Modellantaganden

Modellen som vi använde för att ta fram skattningarna bygger på ett antal antaganden:

- Det finns ett linjärt samband mellan x och y.
- Mätfelen/slumpavvikelserna ϵ_i är normalfördelade och har alla samma varians σ^2.

Hur kan vi undersöka dessa antaganden?

Residualstudier

Låt

$$e_i = y_i - (\hat{m} + \hat{k}x_i)$$

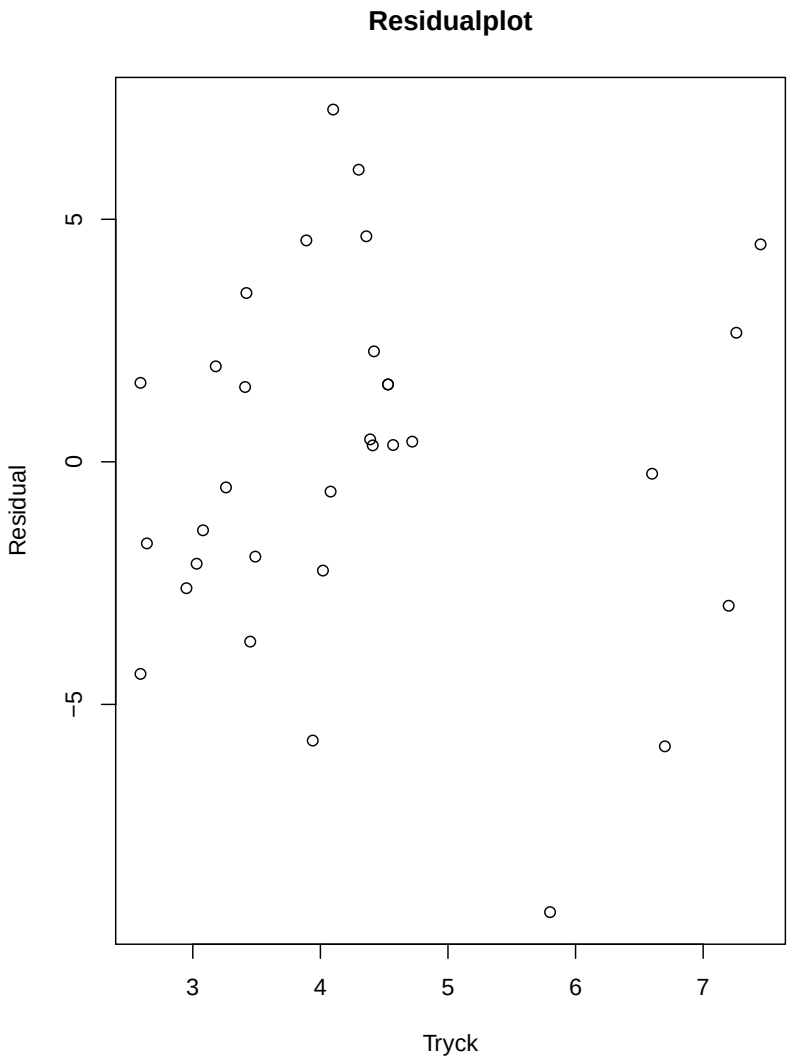
Residualerna e_i beskriver datamaterialets spridning kring den anpassade linjen.

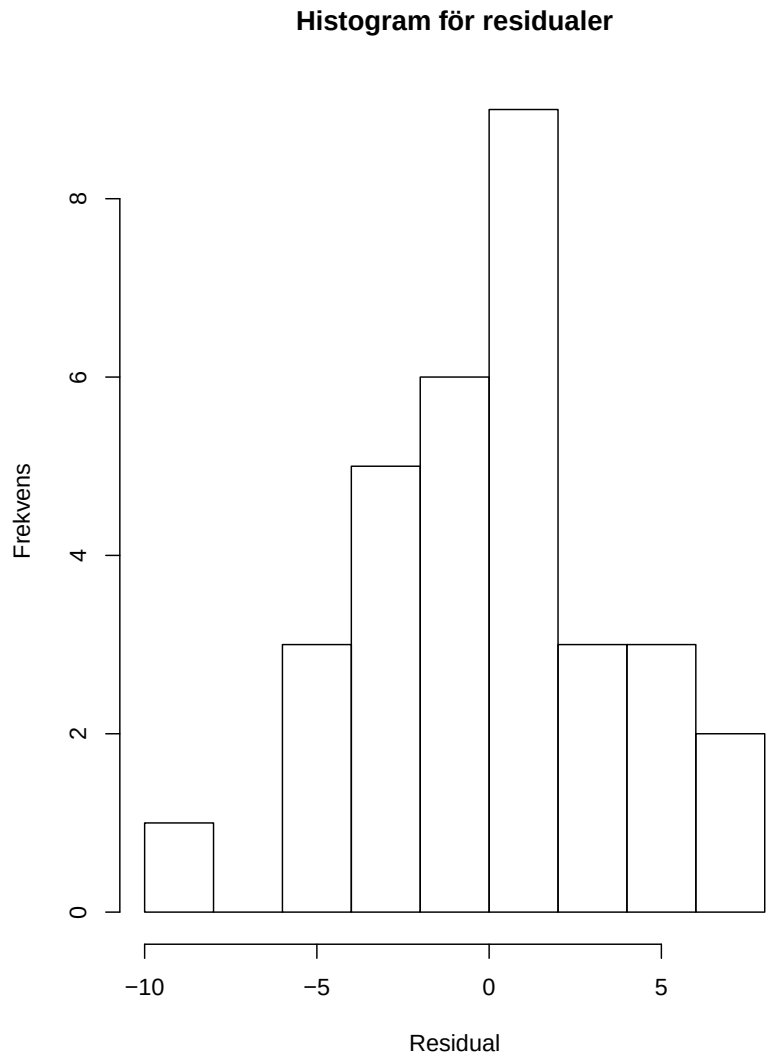
- Se tavlan!

Genom att plotta residualerna och exempelvis rita deras histogram så kan antagandet om oberoende mätfel och att $\epsilon_i \sim N(0, \sigma^2)$ undersökas.

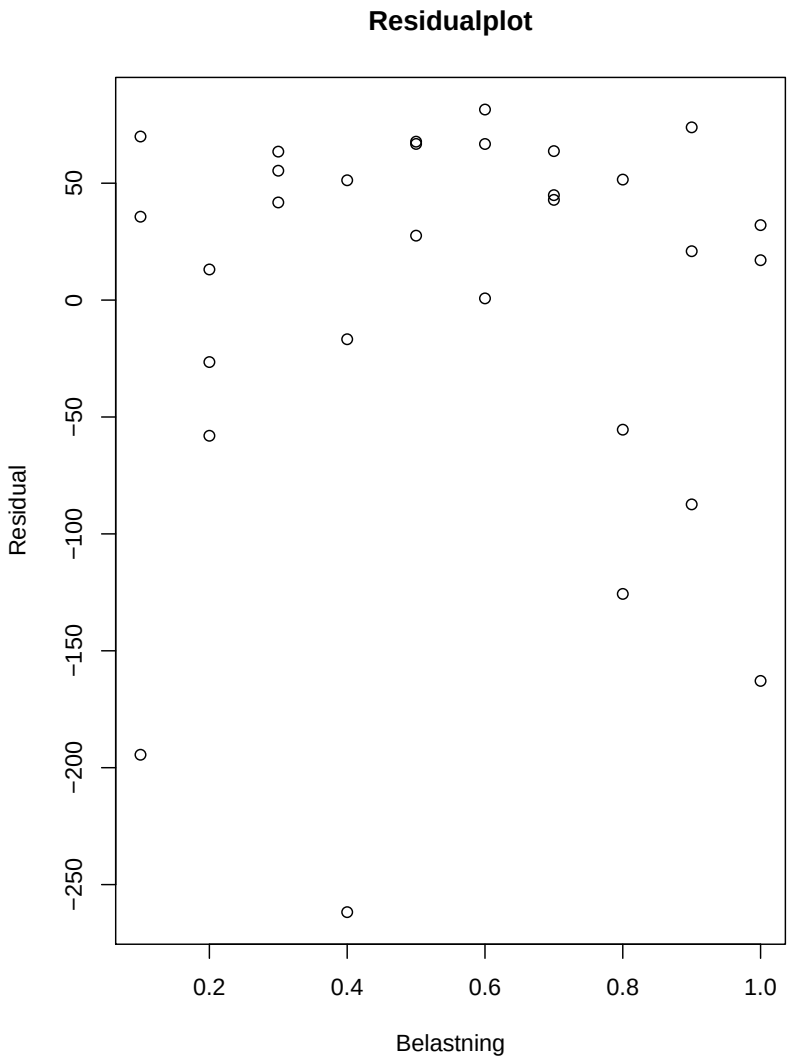
- Mönster i residualplotten tyder på beroende mätfel eller att modellen är felaktig.
- Skillnader i residualernas spridning för olika x_i tyder på att variansen σ^2 inte är konstant utan beror på x_i.
- Histogram som inte liknar normalfördelningens täthetsfunktion tyder på att mätfelen inte är normalfördelade.

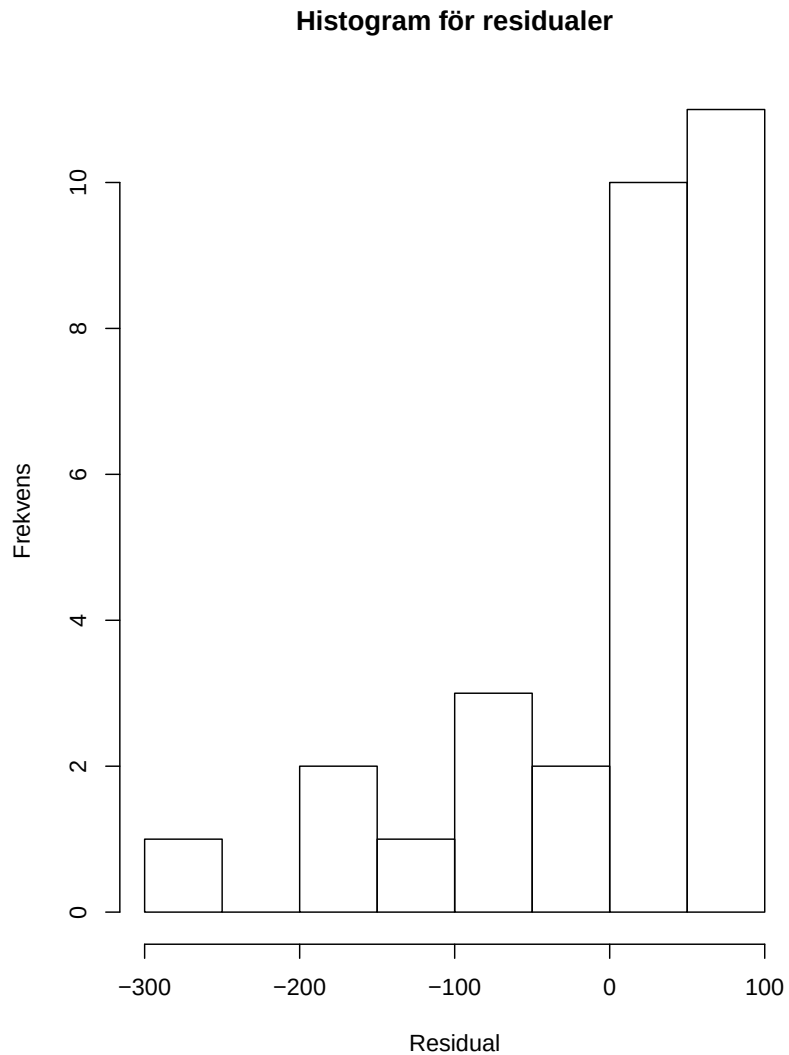
Residualstudier: tryck i bensintank





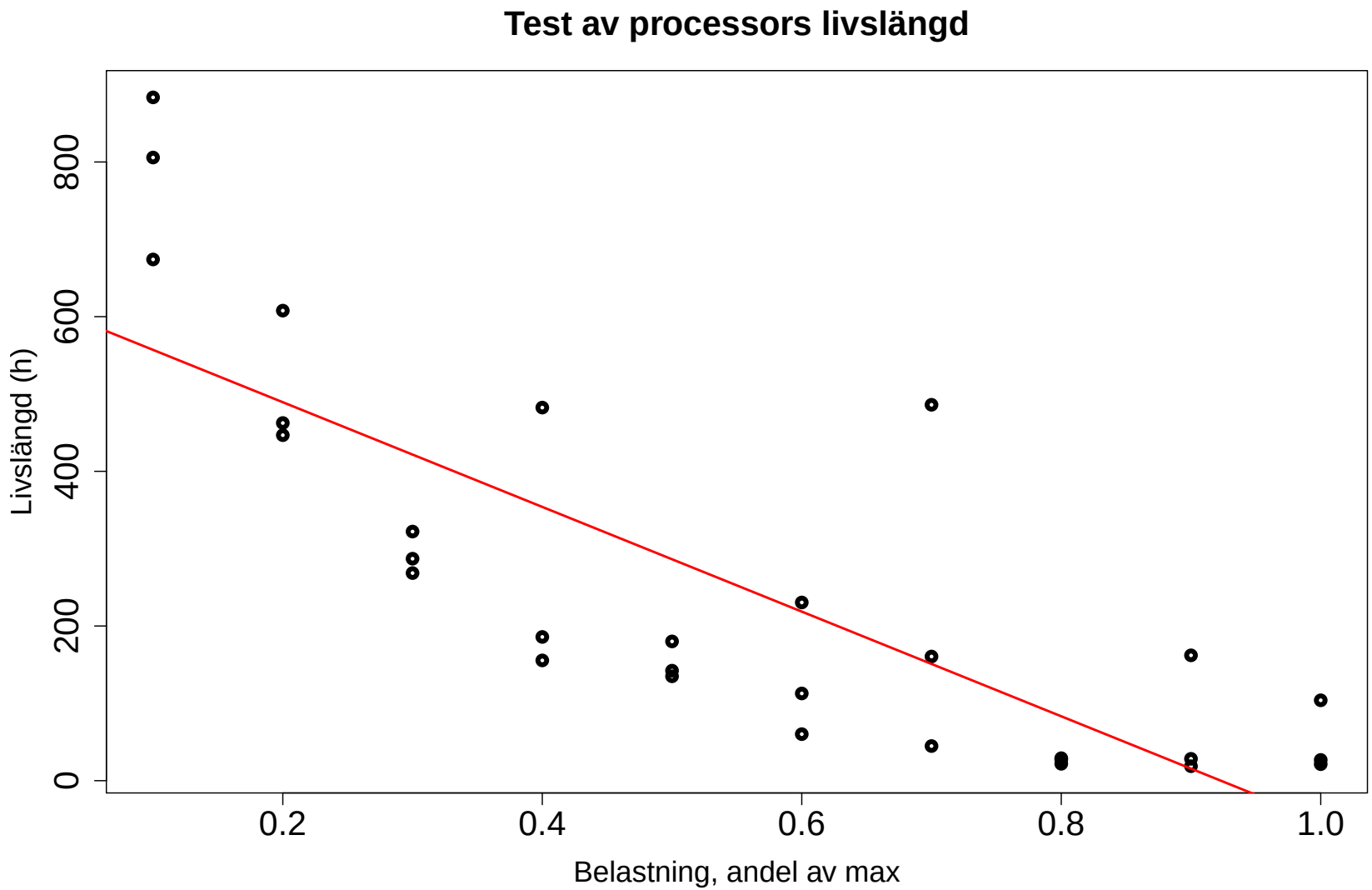
Residualstudier: processorbelastning



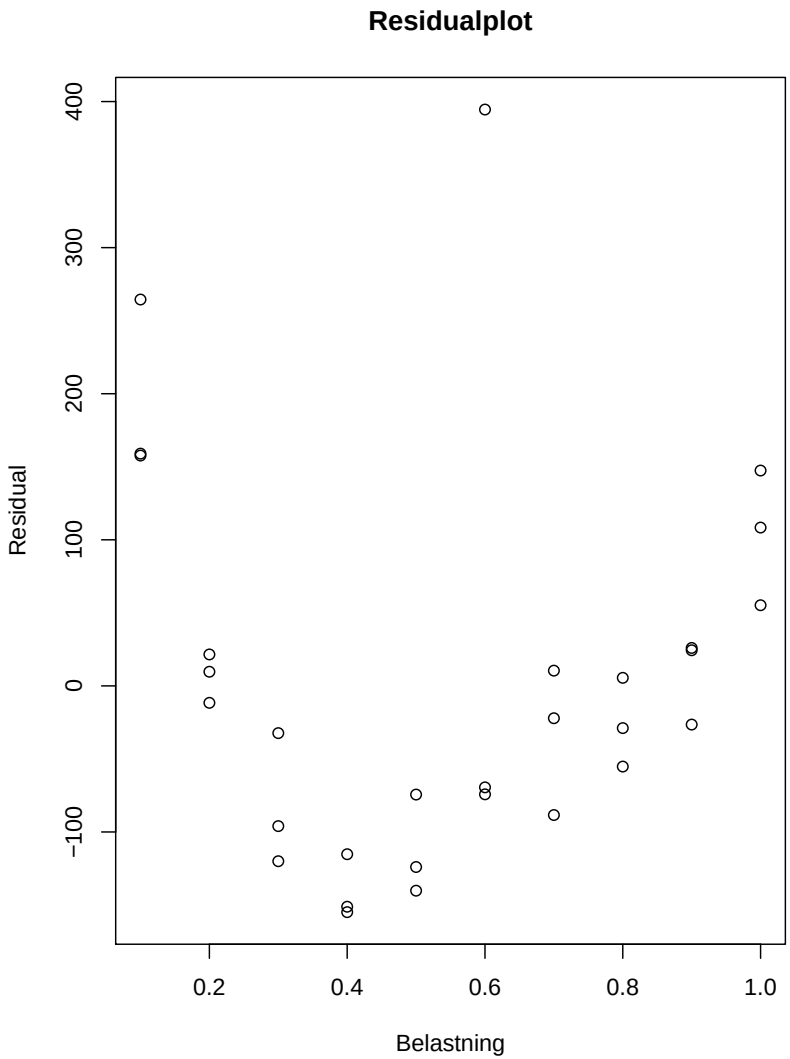


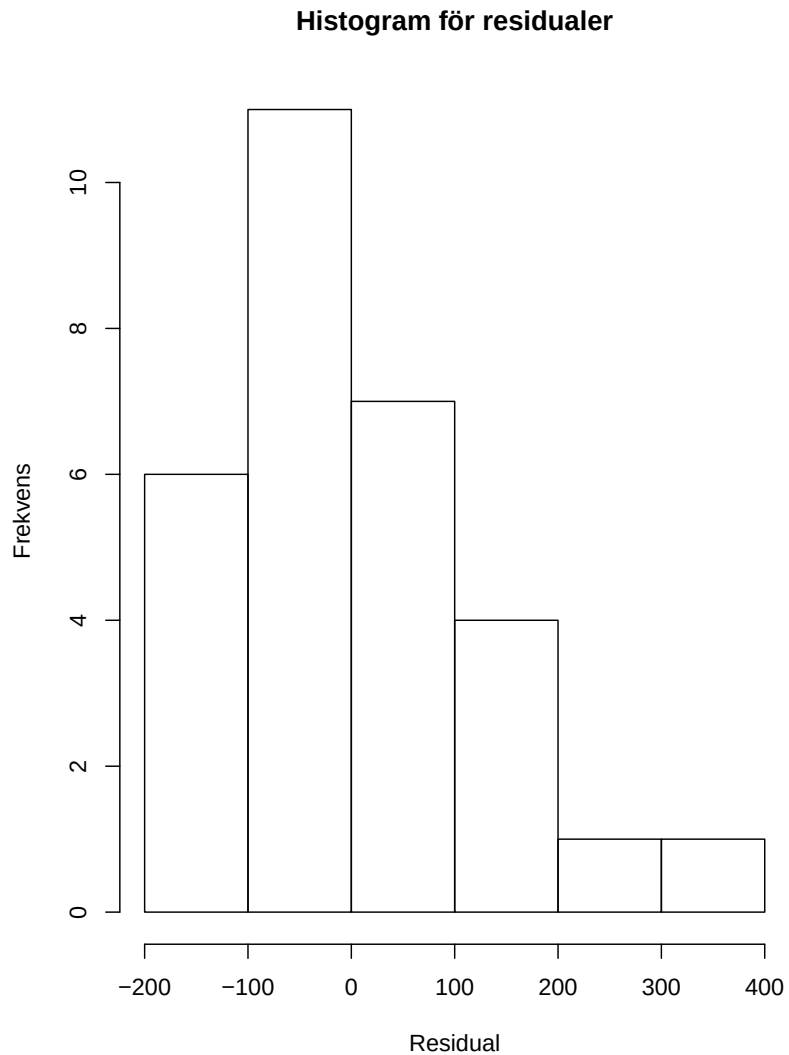
Residualstudier: processorbelastning

Anpassning av en rät linje till ursprungliga (icke-logaritmerade) processordata:



Residualstudier: processorbelastning





A-delen från tentan 2011-03-17

(h) För observationerna $(x_1, y_1), \ldots, (x_{15}, y_{15})$ har man anpassat modellen

$$y = \alpha + \beta x + \epsilon, \quad \text{där } \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Statistisk programvara gav punktskattningarna $\hat{\alpha} = -1.9$, $\hat{\beta} = -0.5$, $\hat{\sigma}^2 = 0.55$. Ange utifrån den anpassade modellen det förväntade värdet på y, givet att $x = 5.0$. (1p)

B-delen från tentan 2011-03-17

5. Ett isoleringsmaterials kompressionsvärden undersöktes under olika betingelser på tryck och temperatur. Under ett försök så hölls temperaturen konstant vid 50° C medan man gjorde 10 mätningar av kompressionen y då trycket x hölls vid olika värden mellan 1 och 5 bar.
   Följande beräknades:

$$\sum_{i=1}^{10} x_i = 30, \quad \sum_{i=1}^{10} y_i = 23.6,$$

$$\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 10.26, \quad \sum_{i=1}^{10} (x_i - \bar{x})^2 = 16.30, \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 8.15.$$

   Man antog följande regressionsmodell:

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

   där ϵ_i är oberoende mätfel med fördelning $\mathrm{N}(0, \sigma^2)$.

   (a) Skatta α, β och σ^2. (3p)

   (b) Skatta den förväntade kompressionen vid temperaturen 50° C och trycket 4.5 bar, samt ange ett 95% prediktionsintervall för kompressionen under dessa förhållanden. (3p)

Konfidensintervall

Om problemet går ut på att räkna ut ett konfidensintervall så behöver man veta två saker:

- Vad man vill ha ett konfidensintervall för (t.ex. μ).
- Vilken information man har fått (t.ex. $\bar{x}$ och σ^2).

Bortsett från intervallen vid regression så finns det fyra huvudproblem:

- KI för väntevärdet: μ.
- KI för skillnaden i väntevärden: $\mu_X - \mu_Y$.
- KI för en andel: p.
- KI för skillnaden i andelar: $p_1 - p_2$.

Dessa kan i sin tur delas upp i olika fall. De fall som brukar kunna dyka upp på tentan presenteras på nästa sida.

Konfidensintervall

1. KI för μ, σ känd: $\quad \bar{x} \pm \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

2. KI för μ, σ okänd: $\quad \bar{x} \pm t_{\alpha/2}^{(n-1)} \cdot \frac{s}{\sqrt{n}}$

3. KI för μ, X ej normalfördelad: om n är tillräckligt stort så kan intervallet $\quad \bar{x} \pm \lambda_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ användas som approximation.

4. KI för $\mu_X - \mu_Y$, σ okänd:
$$\bar{x} - \bar{y} \pm t_{\alpha/2}^{(n_x+n_y-2)} \cdot s_p \sqrt{1/n_x + 1/n_y}$$

5. KI för $\mu_X - \mu_Y$, stickprov i par: ta differenserna $z_i = x_i - y_i$ och använd $z_1, \ldots, z_n$ i intervallen 1-2!

6. KI för p: $\quad \hat{p} \pm \lambda_{\alpha/2} \sqrt{\frac{1}{n}\hat{p}(1-\hat{p})} \quad$ om $\hat{p}(1-\hat{p})n \geq 10$.

7. KI för $p_1 - p_2$:
$$\hat{p}_1 - \hat{p}_2 \pm \lambda_{\alpha/2} \sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)}$$

B-delen från tentan 2012-03-14

3. I en studie jämfördes smältpunkten för två legeringar som används som lod vid lödning genom att man smälte 20 stycken prov av vardera legering och mätte temperaturerna. För den första legeringen fick man $\bar{x}_1 = 215$ och $s_1 = 2.2$ och för den andra legeringen fick man $\bar{x}_2 = 218$ och $s_2 = 2.8$. De uppmätta smältpunkterna kan ses vara normalfördelade. Kan man utifrån detta säga att den andra legeringen har en högre smältpunkt än den första?

(6p)