Løsningsforslag til Re-eksamen i M4STI1 2015F

Opgave 1

En robot til sprøjtemaling laver utilsigtede 'helligdage', d.v.s. små pletter, der ikke er blevet dækket af maling. Robotten laver i gennemsnit 0.8 helligdage per malet kvadratmeter. Robotten skal male 70 cirkelformede skiver på forsiden. Hver skive har en diameter på 1.2 m.

a. *Hvor mange helligdage må der forventes at være på en tilfældig skive?*
   Arealet af en skive er $A = \pi(\frac{1}{2} \cdot 1.2\ \mathrm{m})^2 = 1.1310\ \mathrm{m}^2$.
   Der må forventes $0.8\ \mathrm{m}^{-2}\ \cdot 1.1310\ \mathrm{m}^2 = \mathbf{0.9048}$ helligdage per skive

b. *Hvilken sandsynlighedsfordeling vil du bruge til at beskrive antal helligdage på en skive, og hvad er fordelingens middelværdi, varians og spredning?*
   Poissonfordelingens middelværdi og varians er begge λ, mens spredningen er $\sqrt{\lambda}$. I dette tilfælde er $\lambda = 0.9048$, så middelværdi og varians er **0.9048**, mens spredningen er **0.9512**

c. *Hvad er sandsynligheden for, at ingen af de 70 skiver har helligdage?*
   Sandsynligheden for at 1 skive har 0 helligdage er p_0_1 = poisspdf(0, 0.9048) = 0.4046.
   Da antal helligdage på skiverne er uafhængige af hinanden, er sandsynligheden for at alle 70 skiver har 0 helligdage $\mathrm{p\_0\_70} = (0.4046)^{70} =$ **3.1201e-28**
   Alternativt kan dette udregnes som p_0_70 = poisspdf(0, 70·0.9048) = 3.1201e-28

d. *Beregn det forventede antal skiver med henholdsvis 0, 1, 2, 3 og 4 eller flere helligdage.*
   Først regnes sandsynligheden ud for at en skive har hhv. 0, 1, 2, 3 og 4 eller flere helligdage. For x = 0, 1, 2 og 3 regnes dette ud som p(x) = poisspdf(x, 0.9048). Sandsynligheden for 4 eller flere helligdage er den resterende sandsynlighedsmasse, altså $(x \geq 4) = 1 - \sum_{x=0}^{3} p(x)$. Vi får:
   p = [0.4046 0.3661 0.1656 0.0499 0.0137]
   Den forventede hyppighed fås ved at gange med antal skiver, 70:
   E = [28.3242 25.6271 11.5934 3.4965 0.9587]

e. *Lav en Goodness of Fit test for, om de observerede og de forventede hyppigheder stemmer overens på 1 % signifikansniveau.*
   I Goodness of Fit testen skal vi sammenligne den forventede hyppighed E med den observerede O, hvor:
   O = [22 31 10 7 0]
   For at være stabil kræver metoden, at det forventede antal er mindst 3 i alle kategorier. Da den sidste kategori (4 eller flere) kun har et forventet antal helligdage på 0.9587 slår vi denne sammen med den næstsidste, så kategorierne er 0, 1, 2 og 3 eller flere:

E = [28.3242 25.6271 11.5934 4.4552]
O = [22 31 10 7]

Hypotesetestens 5 skridt:

1. Hypoteser:
H0: Antal helligdage på en skive følger en poissonfordeling med $\lambda = 0.9048$
H1: Antal helligdage på en skive følger *ikke* denne fordeling.

2. Teststatistik:
$c = \sum_{i=0}^{k-1} \frac{(O_i - E_i)^2}{E_i}$ er chi-i-anden fordelt med df frihedsgrader, hvor df = k-p-1; k er antal kategorier (4), p er antal parametre til fordelingen, som vi har estimeret fra observationerne (p=0, for vi kendte $\lambda = 0.9048$). Derfor er df = 4 - 0 - 1 = 3.

3. Kritisk grænse:
Signifikansniveauet er $\alpha = 0.01$, så den kritiske grænse er den værdi, hvor kun 1% af sandsynlighedsmassen ligger over. Vi beregner den kritiske grænse i MatLab som:
chi2_0 = chi2inv(1-alpha, df)
Vi får chi2_0 = 11.3449.

4. Beregn teststatistik:
Teststatistikken beregnes i MatLab som:
chi2test = sum(((O - E).^2)./E)
Vi får chi2test = 4.2111.

5. Konklusion
Da teststørrelsen ikke er over den kritiske grænse kan vi ikke forkaste nulhypotesen. Antal helligdage følger altså en poissonfordeling med $\lambda = 0.9048$

Opgave 2

Sammenhørende værdier for vindhastighed (x) og den strøm, en bestemt vindmølle producerer (y):

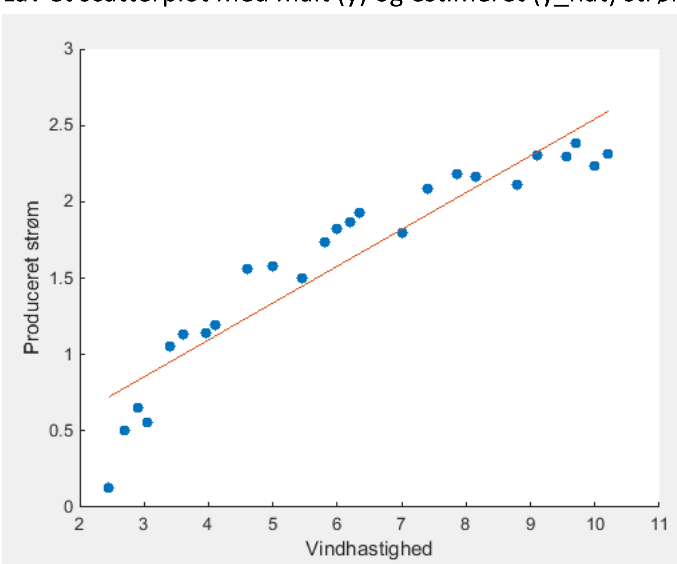
a. *Lav en lineær regressionsanalyse af produceret vindmøllestrøm som funktion af vindhastighed og skriv regressionsligningen op. Beregn den forventede produktion ved en vindhastighed på 9.5.*

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    ________    ______    __________

    (Intercept)    0.13088      0.12599    1.0388       0.30971
    x1             0.24115     0.019049    12.659    7.5455e-12


Number of observations: 25, Error degrees of freedom: 23
Root Mean Squared Error: 0.236
R-squared: 0.874,  Adjusted R-Squared 0.869
F-statistic vs. constant model: 160, p-value = 7.55e-12
```

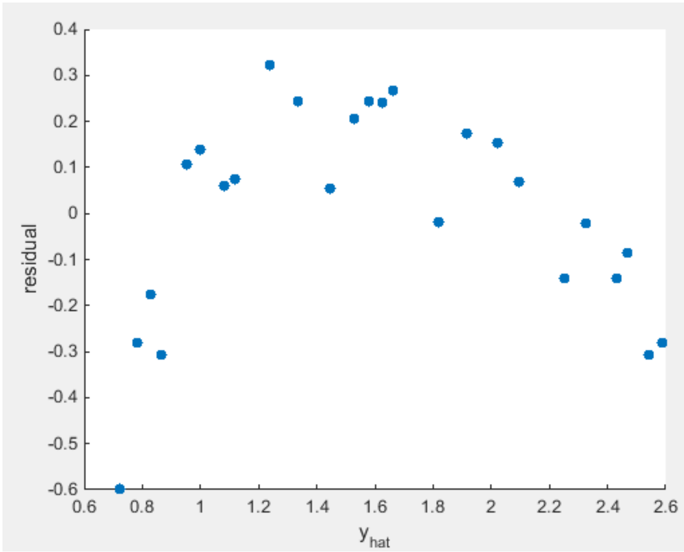
y = 0.13088 + 0.24115*x
y_95 = 0.13088 + 0.24115*9.5 = 2.4218
Forventet strømproduktion svarende til 9.5 er 2.4218

b. Forklar v.h.a. regressionsanalysens statistikker (f.eks. R-squared og p-value), om modellen beskriver observationerne godt.

c. Lav et scatterplot med målt (y) og estimeret (y_hat) strømproduktion som funktion af vindhastighed.



Lav desuden et residualplot (residual mod y_hat). Hvad viser de to plots om regressionsmodellen?



d. Forsøg at forbedre modellen med transformationer. Prøv følgende to modeller (henholdsvis en logaritmisk og en reciprok transformation af vindhastigheden):

$$y = b_0 + b_1 \ln(x)$$

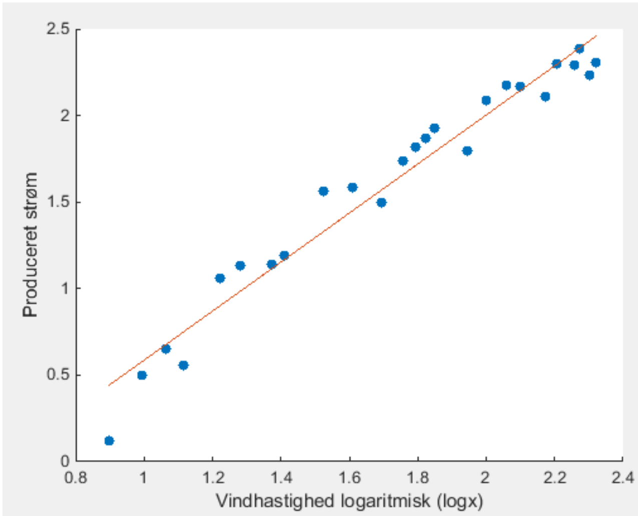
$$y = b_0 + b_1 \, {}^{1}\!/_{x}$$

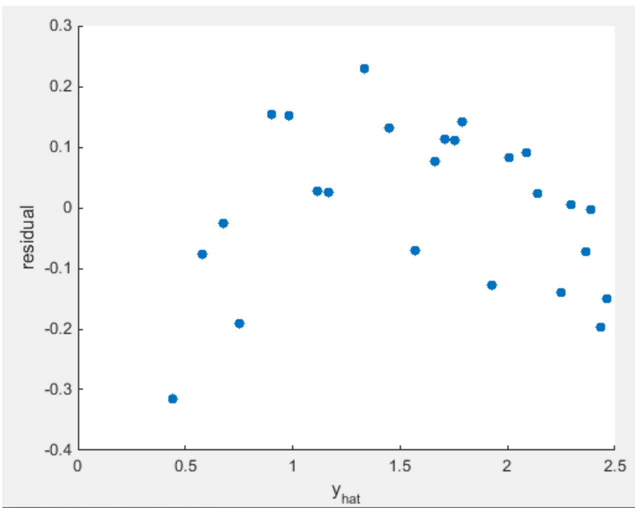
Skriv funktionsudtrykkene for de to transformerede modeller op.

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)     2.9789     0.044902     66.341    8.9177e-28
    x1             -6.9345      0.20643    -33.592    4.7425e-21


Number of observations: 25, Error degrees of freedom: 23
Root Mean Squared Error: 0.0942
R-squared: 0.98,  Adjusted R-Squared 0.979
F-statistic vs. constant model: 1.13e+03, p-value = 4.74e-21
```

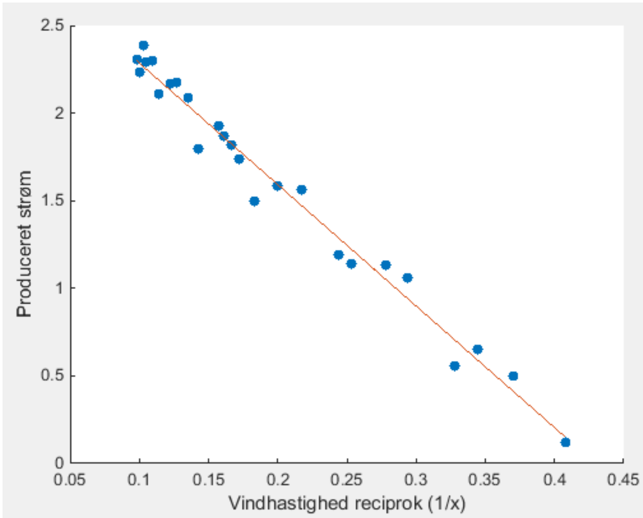

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)     2.9789     0.044902     66.341    8.9177e-28
    x1             -6.9345      0.20643    -33.592    4.7425e-21


Number of observations: 25, Error degrees of freedom: 23
Root Mean Squared Error: 0.0942
R-squared: 0.98,  Adjusted R-Squared 0.979
F-statistic vs. constant model: 1.13e+03, p-value = 4.74e-21
```

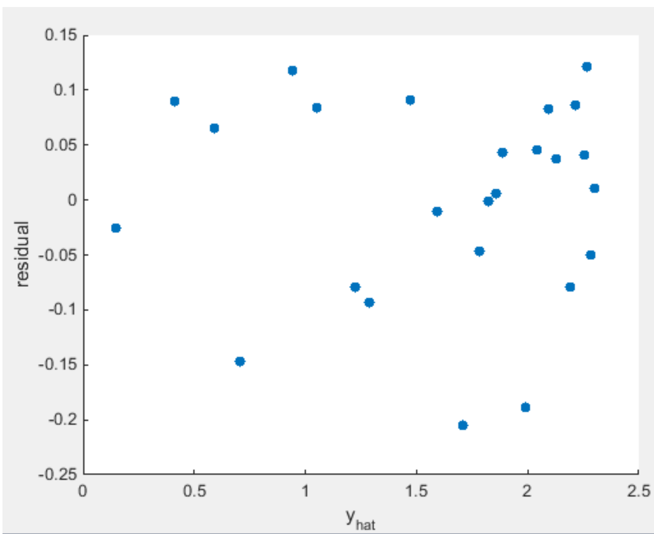
e. Lav scatterplots og residualplots af de to transformerede modeller.
Logaritmisk transformation:





Reciprok transformation:





f. Diskutter hvilken model, der er bedst.

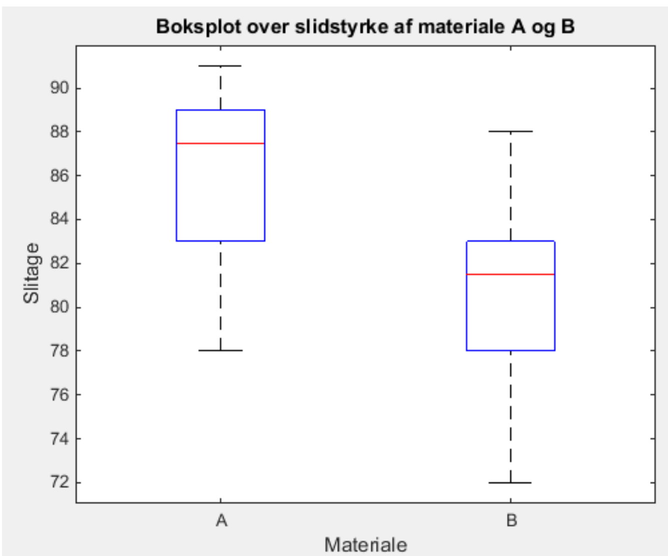
Opgave 3

En særlig maskine bruges til at måle slitage på emner på en standardiseret måde. Det foregår ved, at emnet spændes fast i maskinen, som dernæst påvirker emnet på en ensartet måde i et fastlagt tidsrum. Til sidst bruger maskinen laserbelysning til at måle den slitage, som emnet er blevet påført. Slitagen måles på et indeks fra 0 til 100, hvor et højt indeks betyder meget slitage.

I et eksperiment blev to forskellige materialer, A og B, testet i maskinen for at måle forskel i materialernes slidstyrke. 12 emner af materiale A og 10 emner af materiale B blev testet. Resultatet vises i følgende tabel:

Materiale	Slitageindeks											
A	89	90	88	91	89	85	87	83	89	78	80	83
B	78	88	83	77	88	72	80	80	83	83		

a. Lav og kommenter et parallelt boksplot, der viser slitageindeks for de to materialer.



b. Man ønsker at slå fast med et signifikansniveau på 5 %, om materiale B er mindst 2 enheder på slitageindekset mere slidstærkt end materiale A. Opstil nulhypotese og alternativhypotese for denne hypotesetest.

c. Opstil og beregn teststatistikken. Angiv hvilken fordeling den følger.

d. Beregn den kritiske region for testen og konkludér på hypotesetesten.

e. Er der forskel på slidstyrken af de to materialer på 5 % signifikansniveau?

f. Beregn et 95 % konfidensinterval for forskellen på materialernes middelværdi.

g. Diskutter hvordan boksplot, hypotesetest og konfidensinterval stemmer overens.

h. Oplys hvilke antagelser, der er gjort i hypotesetesten, og om antagelserne er rimelige på baggrund af data.