Sannolikhetslära och Statistik
Föreläsning 11

Mattias Villani

Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Översikt

- Enkel regression
- Kovarians och korrelation
- Multipel regression
- Regression med binär respons

Regression

- Hittills: modeller utan förklaringsvärde. $\theta=$ sannolikheten för spam.
- Samma spam-sannolikhet, θ, för:
  - ett mejl med 256 $-tecken, som inte nämner mitt namn, och som kommer från avsändare utanför min adressbok
  - ett mejl utan $-tecken, som nämner mitt namn, och som kommer från en avsändare i min adressbok.
- Prediktion av siffror: vi vill koppla klassen (0-9) till gråheten i pixlarna.
- Lösning: låt θ vara en funktion av förklaringsvariabler, t ex antal$, mittNamn, kändAvsändare etc.
- **Regression**: låt fördelning för en **responsvariabel** Y (t ex binära Spam/Ham) bero på ett antal **förklarande variabler** $X^{(1)}, ..., X^{(k)}$ (alt. namn: prediktorer, kovariater, oberoende variabler).

Enkel regression

- **Enkel regression**: en enda förklarande variabel, X. X antas känd (ej stokastisk).
- Regression modellerar den betingade fördelningen $f(Y|X = x)$.
- Vanligast: X påverkar bara väntevärdet i fördelningen: $E(Y|X = x)$.
- Antag $Y|(X = x) \sim N\left(\mu\left(x\right), \sigma^2\right)$, där

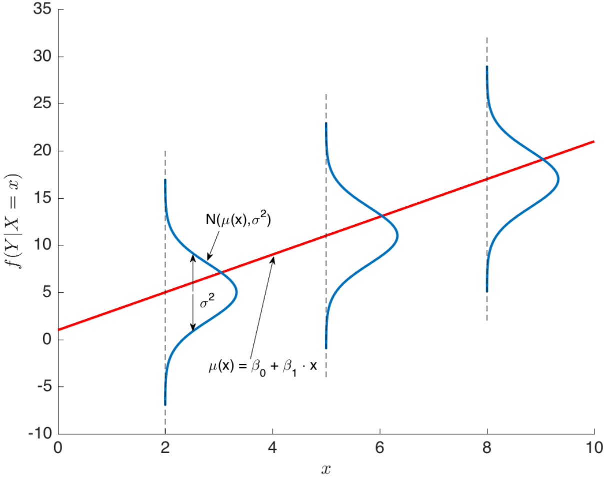
$$E(Y|X = x) = \mu\left(x\right) = \beta_0 + \beta_1 x$$

- Kan också skrivas

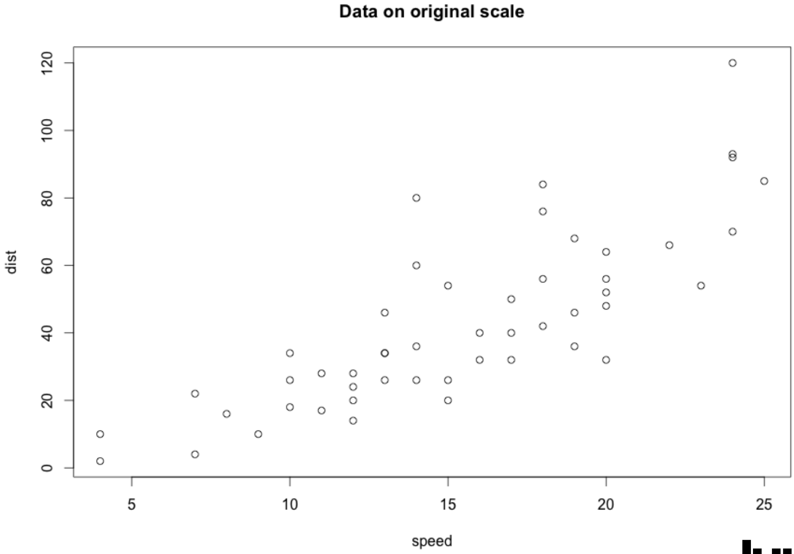
$$Y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

- ε kallas för **störning** eller **felterm**.

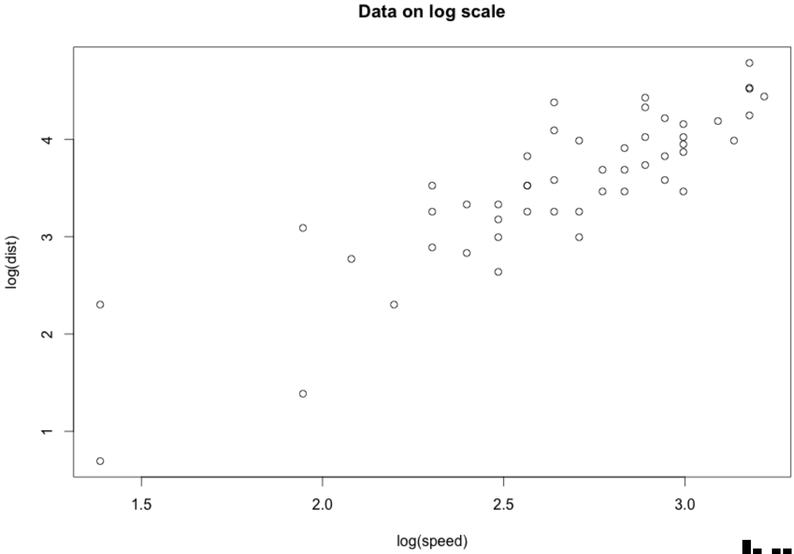
Enkel regression



Exempel: stoppsträcka=f(hastighet)



EXEMPEL: STOPPSTRÄCKA=F(HASTIGHET)



Estimation - minsta kvadratmetoden

- **Data** är X-Y talpar: $(x_1, y_1), ..., (x_n, y_n)$.
- **Regressionlinjen** $\beta_0 + \beta_1 \cdot x$ ger prognoserna:
  $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i,\ i = 1, ..., n.$
- **Residualen** vid x_i

$$e_i = y_i - \hat{y}_i$$

- Minsta kvadratmetoden: välj β_0 och β_1 så summan av kvadrerade residualerna minimeras

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i)^2$$

  - (partial)derivera med avseende på β_0 och β_1 och lös ekvationssystemet

$$\frac{\partial Q}{\partial \beta_0} = 0$$
$$\frac{\partial Q}{\partial \beta_1} = 0$$

Estimation - minsta kvadratmetoden

- (partial)derivera med avseende på β_0 och β_1 och lös ekvationssystemet

$$\frac{\partial Q}{\partial \beta_0} = 0$$
$$\frac{\partial Q}{\partial \beta_1} = 0$$

ger lösningen

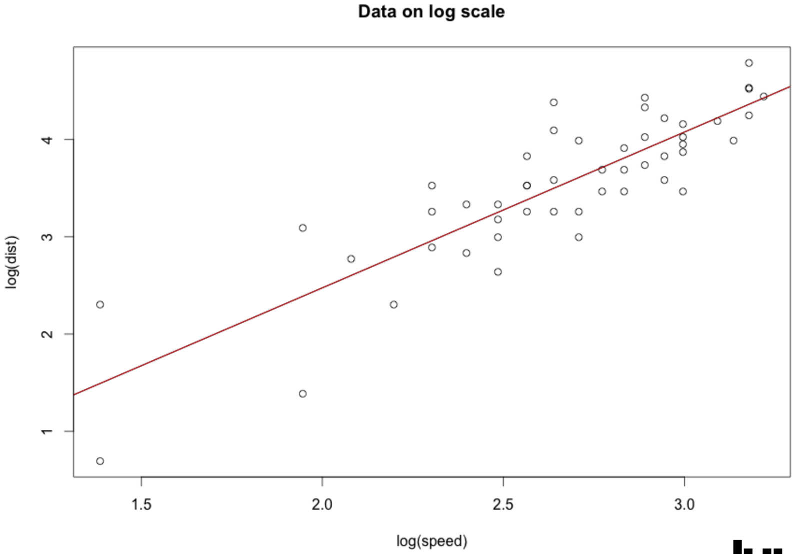
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

REGRESSION I R

```r
data(cars) # Loading one of R's internal data sets
attach(cars) # Making variables in cars available (outside of 'namespace')
lmFit <- lm(log(dist) ~ log(speed)) # general:lm(y ~ x1 + x2 + x1*x2)
summary(lmFit)
```

```
##
## Call:
## lm(formula = log(dist) ~ log(speed))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00215 -0.24578 -0.02898  0.20717  0.88289
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7297     0.3758  -1.941   0.0581 .
## log(speed)    1.6024     0.1395  11.484 2.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4053 on 48 degrees of freedom
## Multiple R-squared:  0.7331,Adjusted R-squared:  0.7276
## F-statistic: 131.9 on 1 and 48 DF,  p-value: 2.259e-15
```

Exempel: stoppsträcka=f(hastighet)



Estimation - maximum likelihood

- **ML**: välj värden på β_0 och β_1 som maximerar sannolikheten (tätheten) för data. Antag oberoende normalfördelade feltermer $(\varepsilon_1, ..., \varepsilon_n)$.
- **Likelihoodfunktionen**

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} N\left(y_i | \mu(x_i), \sigma^2\right)$$

där $N\left(y_i | \mu(x_i), \sigma^2\right)$ är tätheten för en $N(\mu(x_i), \sigma^2)$ fördelning

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \mu(x_i)\right)^2\right)$$

Alltså

$$L(\beta_0, \beta_1) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2\right)$$

Estimation - maximum likelihood

- **Likelihoodfunktionen**

$$L(\beta_0, \beta_1) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2\right)$$

- Vi kan lika gärna maximera **log-likelihoodfunktionen**

$$\ln L(\beta_0, \beta_1) = c - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2,$$

där $c = -n\ln\left(\sqrt{2\pi\sigma^2}\right)$ är en konstant som inte beror på β_0 och β_1.

- Maximera $\ln L(\beta_0, \beta_1)$ är detsamma som minimera $\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2$.
- **ML = minsta kvadrat!**

Regression och korrelation

- **Kovarians**

$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right]$$

- **Korrelationskoefficient** $-1 \leq \rho \leq 1$

$$\rho = \frac{Cov(X, Y)}{Std(X) \cdot Std(Y)}$$

- **Stickprovskovarians** (väntevärdesriktig estimator av $Cov(X, Y)$):

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- **Stickprovskorrelationskoefficient**

$$r = \frac{s_{xy}}{s_x s_y}, \text{ där } s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$$

- Relation mellan regression och korrelation

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2} = r \cdot \frac{s_y}{s_x}.$$

Multipel regression

- Fler än en förklarande variabel.
- Antag

$$Y|X^{(1)} = x^{(1)}, ..., X^{(k)} = x^{(k)}) \sim N\left(\mu\left(x^{(1)}, ..., x^{(k)}\right), \sigma^2\right)$$

där

$$\mu\left(x^{(1)}, ..., x^{(k)}\right) = \beta_0 + \beta_1 x^{(1)} + ... + \beta_k x^{(k)}$$

- Kan också skrivas

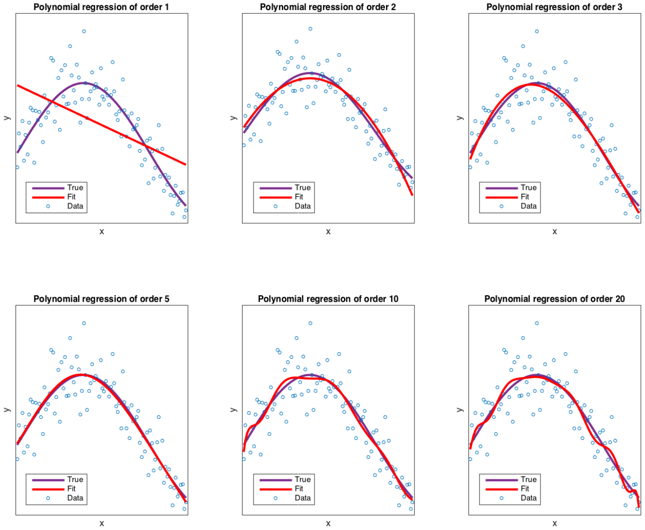
$$y = \beta_0 + \beta_1 x^{(1)} + ... + \beta_k x^{(k)} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

- Se Baron 11.3.2 för minsta kvadrat. ML = minsta kvadrat.
- **Polynomregression** för icke-linjär regression

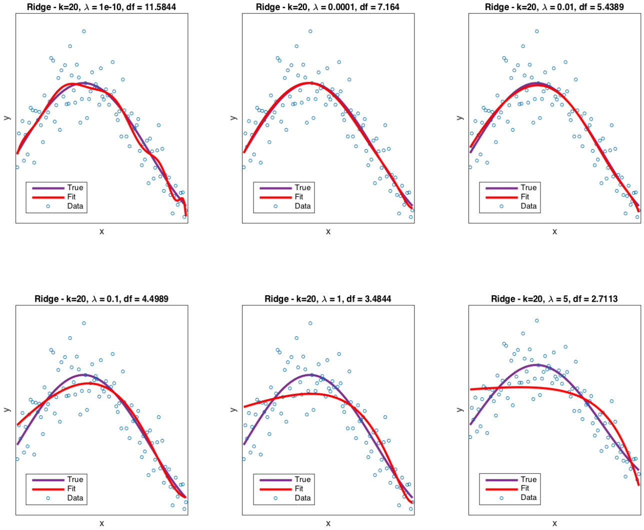
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$$

- Kan skattas med minsta kvadrat. Se upp för **överanpassning**!

Överanpassning



Överanpassning - mjukhetsprior



Regression med binär respons

- Hittills har vi antagit kontinuerlig (normalfördelad) respons, Y.
- Om Y är **binär** kan vi inte anta $Y|(X = x) \sim N(\mu(x), \sigma^2)$.
- Bättre med
$$Y|(X = x) \sim Bernoulli(\theta(x))$$
och olika Y-observationer är oberoende (givet x).
- Vanlig funktionsform för $\theta(x)$ (**logistisk regression**):
$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 \cdot x)}{1 + \exp(\beta_0 + \beta_1 \cdot x)}$$
- Minsta kvadrat är inte längre en bra estimationsmetod.
- **Maximum likelihood** funkar alltid:

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \theta(x_i)^{y_i} (1 - \theta(x_i))^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \beta_1 \cdot x)}{1 + \exp(\beta_0 + \beta_1 \cdot x)} \right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \beta_1 \cdot x)} \right]^{1-y_i} \end{aligned}$$

ML skattningar i logistisk regression

```r
# Defining the log-likelihood function
LogLik <- function(betaVect,y,X){
        linFunc = X%*%betaVect
        thetaVect = exp(linFunc)/(1+exp(linFunc))
        logLikelihood <- sum(y*log(thetaVect) + (1-y)*log(1-thetaVect))
}

# Reading in fraud data from file
data <- read.csv('/Users/matvi05/Dropbox/Teaching/ProbStatUProg/Data/banknoteFraud.csv', header = FALSE)
names(data) <- c("varWave","skewWave","kurtWave","entropyWave","fraud")
y <- data[,5]
X <- as.matrix(cbind(1,data[,1:4]))      # Adding a column of ones for the intercept
nPara <- dim(X)[2]                       # Number of covariates incl intercept

# Optimize to the find the ML estimates.
initPar <- matrix(0,nPara,1)
optimResults <- optim(initPar, LogLik, gr = NULL, y, X, control=list(fnscale=-1))
optimResults$par # betaHat, the ML estimates of beta = (beta0,beta1,...,beta4)


##            [,1]
## [1,]  7.3425752
## [2,] -7.8714117
## [3,] -4.1976080
## [4,] -5.2960804
## [5,] -0.6052862
```