TDAB01 Sannolikhetslära och Statistik

Jose M. Peña
IDA, Linköpings Universitet

Föreläsning 11

Översikt

- **Enkel regression**
- **Estimation: Minsta kvadrat och ML metoderna**
- **Multipel regression**
- **Logistisk regression**

Regression

- Spam/Ham ~ *Bernoulli*(θ).
- Hittills: Modeller utan förklaringsvärde, dvs 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.
- 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** $X^{(1)}, \ldots, X^{(k)}$, också kallade prediktorer, kovariater, oberoende variabler.

Enkel regression

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

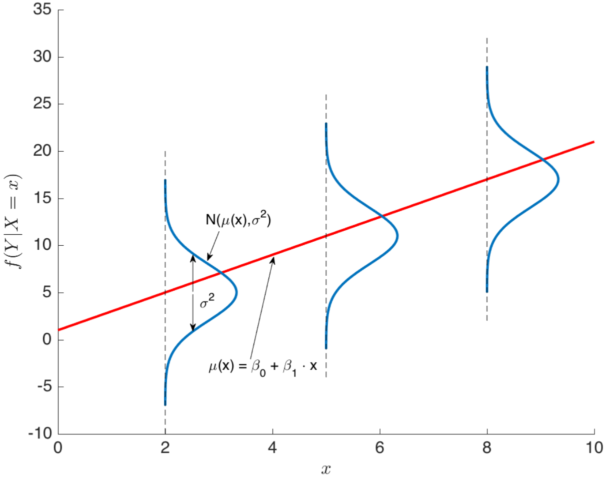
$$E(Y|X=x) = \mu(x) = \beta_0 + \beta_1 x$$

- Kan också skrivas

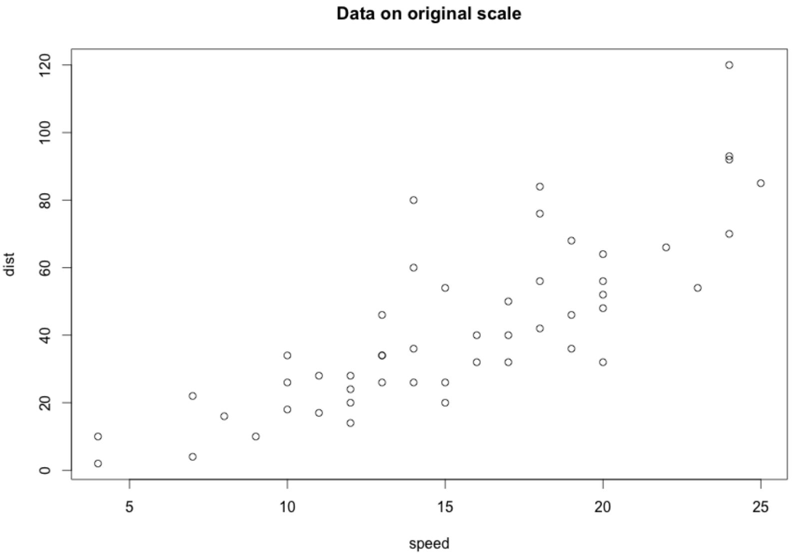
$$Y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

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

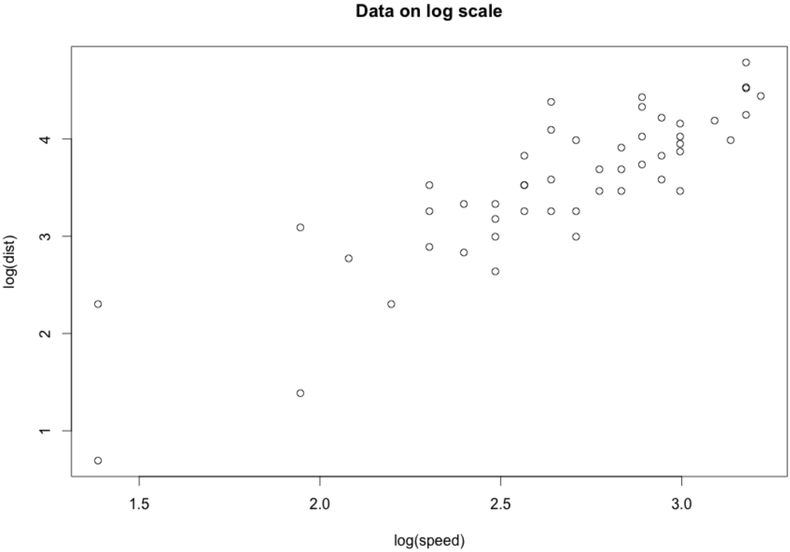
Enkel regression



Exempel: Stoppsträcka som en funktion av hastighet



Exempel: Stoppsträcka som en funktion av hastighet



Estimation: Minsta kvadrat metoden

- Data är X-Y talpar: $(x_1, y_1), \ldots, (x_n, y_n)$.
- **Regressionlinjen** $\beta_0 + \beta_1 x$ ger prognoserna

$$\hat{y}_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- **Residualen** vid x_i:

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

- **Minsta kvadrat metoden**: Välj β_0 och β_1 så summan av kvadrerade residualerna minimeras

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

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

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

som ger lösningen

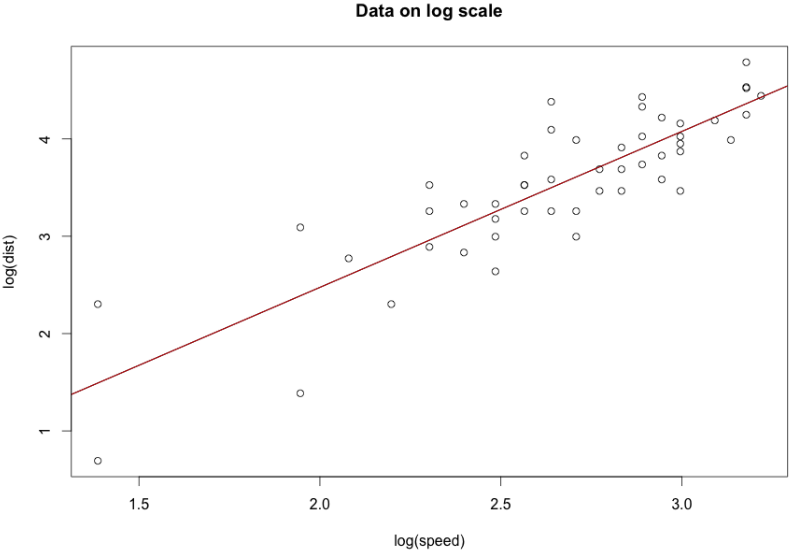
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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 som en funktion av hastighet



Estimation: Maximum likelihood metoden

- ML metoden: Välj värden på β_0 och β_1 som maximerar sannolikheten (tätheten) för data. Antag oberoende normalfördelade feltermer $(\varepsilon_1, \dots, \varepsilon_n)$.
- 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 normal fördelning, dvs

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

  Alltså

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

- 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 !**

Multipel regression

- Fler än en förklarande variabel.
- Antag

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

  där

$$\mu(x^{(1)}, \ldots, x^{(k)}) = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_k x^{(k)}$$

- Kan också skrivas

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

- Minsta kvadrat: $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ där data är

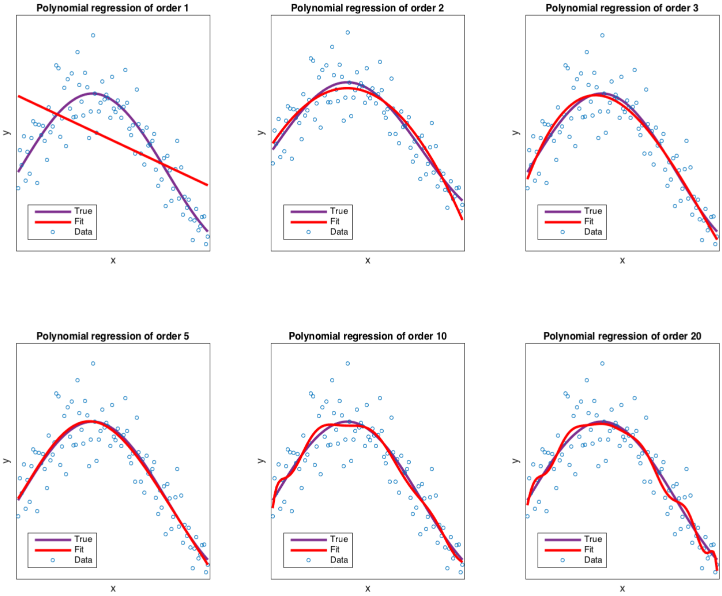
$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_1^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & \ldots & x_n^{(k)} \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

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

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

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

Överanpassning



Logistisk regression

- Hittills har vi antagit kontinuerlig (normalfördelad) respons variabel Y.
- Om Y är **binär** kan vi inte anta $Y|X = x \sim N(\mu(x), \sigma^2)$.
- Istället antar vi
$$Y|X = x \sim \mathit{Bernoulli}(\theta(x))$$
- Vanlig funktionsform för $\theta(x)$: **Logistisk regression**
$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$
som garanterar att $0 \leq \theta(x) \leq 1$.
- Minsta kvadrat är inte längre en bra estimationsmetod.
- ML funkar alltid:
$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \theta(x_i)^{y_i} (1 - \theta(x_i))^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \beta_1 x)} \right]^{1-y_i} \end{aligned}$$
men kan inte lösas analytiskt.

Logistisk regression i R

```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
```

Översikt

- **Enkel regression**
- **Estimation: Minsta kvadrat och ML metoderna**
- **Multipel regression**
- **Logistisk regression**