11 Enkel linjär regression

Man är intresserad av om det finns linjärt samband mellan två variabler, till exempel om temperaturen av havsvattnet kan antas vara en linjär funktion av djupheten. Då har man två variabler, djuphet X och temperatur Y. Man vill beskriva temperaturen i olika (exakta) djuphetsnivåer, dvs. att man fixerar $X = x$ och mäter Y. Då varierar Y fortfarande beroende på andra faktorer. Man har en betingad stokastisk variabel $Y|x$ (Y givet att $X = x$), dess väntevärde betecknas av $\mu_{Y|x}$ och som är en funktion av x.

Grafen av den funktionen ($\mu_{Y|x}$ som en funktion av x) kallas en **regressionskurva** av Y på x. Y kallas en **responsvariabel** eller **beroende variabel** och X en **prediktor** eller **oberoende variable**.

Problemet är nu att skatta $\mu_{Y|x}$ när värdena $x_1, x_2, ..., x_n$ är fixerade. Hur väljer man dessa värden?

1) Man kan fixera $x_1, x_2, ..., x_n$ först (preselected); kontrollerad studie.

2) Man kan inte alltid fixera värdena på förhand och måste använda de värdena som finns; observationsstudie.

I både fall blir stickprovet $(x_1, Y_1|x_1), (x_2, Y_2|x_2), ..., (x_n, Y_n|x_n)$.

11.1 Model och parameterskattning

Man tar bara linjär regression, dvs. att sambandet mellan Y och X är linjärt. Man kan skriva att

$$\mu_{Y|x} = \beta_0 + \beta_1 x, \quad \beta_0, \beta_1 \in \mathbf{R}.$$

För att få en skattning för linjen, måste man skatta β_0 och β_1. Man antar att X är mätt utan fel och att Y är stokastisk. Låt E_i vara skillnaden mellan $Y|x_i$ och dess väntevärde $\mu_{Y|x_i}$, dvs. att

$$E_i = Y|x_i - \mu_{Y|x_i}.$$

Då får man att

$$Y|x_i = \mu_{Y|x_i} + E_i = \beta_0 + \beta_1 x_i + E_i.$$

Man antar att E_i, som är stokastisk, har väntevärde noll. Ofta skriver man bara

$$Y_i = \beta_0 + \beta_1 x_i + E_i.$$

Modellen ovan kallas en **enkel linjär regressionsmodel**.

När man har en regressionsproblem, plottar man först värdena x, y (scattergram). Om punkterna har en linjär trend, är linjär regression lämplig. Sedan skattar man β_0 och β_1 från data. Låt skattningarna vara b_0 resp. b_1. Man kan sedan skriva

$$\hat{\mu}_{Y|x} = b_0 + b_1 x.$$

Punkterna kommer inte att ligga exakt på linjen och därför skriver man

$$y_i = b_0 + b_1 x_i + e_i,$$

där e_i kallas **residualen** (avståndet mellan punkten och den skattade regressionslinjen).

Man kan använda en så-kallad **minstakvadratsmetoden** (method of least squares) för att skatta β_0 och β_1. Man vill hitta den linjen som passar bäst för data: Man väljer b_0 och b_1 sådana att de minimerar kvadratsumman av residualerna

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

b_0 och b_1 hittar man genom att derivera SSE m.a.p. b_0 och b_1. Man får att

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

och

$$b_0 = \bar{y} - b_1 \bar{x}.$$

Ex.11.1.1: Fuktigheten påverkar evaporation och därför påverkas lösningsbalansen av vattenreducerbara färg av fuktigheten under sprutning (målning). Man studerar sambandet mellan fuktigheten X och omfattningen av evaporation Y. Studien hjälper målarna att justera sina verktyg. (Data: $n = 25$, $\sum x_i y_i = 11824.44$, $\sum x_i = 1314.90$, $\sum y_i = 235.70$ och $\sum x_i^2 = 76308.53$).

11.2 Egenskaper av skattningar $\hat{\beta}_0$ och $\hat{\beta}_1$

Den linjära regressionsmodellen kan skrivas

$$Y_i = \beta_0 + \beta_1 x_i + E_i,$$

där E_i är en stokastisk variabel med väntevärde noll. För att kunna säga någonting av egenskaperna av $B_0 = \hat{\beta}_0$ och $B_1 = \hat{\beta}_1$, måste man anta någon fördelning för E_i. Vanligen antar man att $E_1, E_2, ..., E_n$ är ett stickprov från normalfördelningen med väntevärdet noll och variansen σ^2. Då är Y_i-erna oberoende och normalfördelade

med resp. väntevärden $\beta_0 + \beta_1 x_i$ och variansen σ^2. Dvs. att Y_i-erna kan ha olika väntevärden men variansen är samma för varje Y_i.

Man kan visa att B_0 och B_1 är normalfördelade

$$B_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}\right)$$

och

$$B_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}\right).$$

Genom att använda dessa fördelningar, kan man intervallskatta B_0 och B_1 och testa hypoteser angående dem.

Ofta känner man inte variansen σ^2 och den måste skattas. Man använder

$$\hat{\sigma^2} = S^2 = \frac{\sum(Y_i - (B_0 + B_1 x_i))^2}{n-2}$$

som en väntevärdesriktig skattning för σ^2.

Nu är t.ex. statistikan

$$\frac{B_1 - \beta_1}{S/\sqrt{\sum(x_i - \bar{x})^2}}$$

T_{n-2}-fördelad och den kan användas för att hitta konfidensintervall för β_1 och testa hypoteser angående β_1. Man kan till exempel testa nollhypotesen $H_0 : \beta_1 = 0$ mot ett av alternativen $H_1 : \beta_1 > 0$, $H_1 : \beta_1 < 0$ eller $H_1 : \beta_1 \neq 0$.

11.5 Residualanalys

Residualen

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y}_i$$

är skillnaden mellan observationen y_i och dess skattade värde. Man kan använda residualer för att säga om den linjära regressionsmodellen är en lämplig model och för att kolla att modelantaganden gäller. När man plottar (x_i, e_i), får man en residualgraf. Den visar om

1) väntevärdet av E_i är noll (punkterna borde vara omkring $e_i = 0$).

2) variansen är konstant (spridningen lika stor med alla värden av x)

3) den linjära modellen inte är bra

4) det finns hål i datamängden.

11.6 Korrelation

Både X och Y är stokastiska. Finns det ett linjärt samband mellan dem?

Korrelationen mellan X och Y definieras

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}},$$

där $Cov(X,Y) = \mathbf{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbf{E}[XY] - \mu_X\mu_Y$.

För att skatta ρ använder man följande skattningar

$$\widehat{Var(X)} = \frac{1}{n}\sum(X_i - \bar{X})^2 =: \frac{S_{xx}}{n},$$

$$\widehat{Var(Y)} = \frac{1}{n}\sum(Y_i - \bar{Y})^2 =: \frac{S_{yy}}{n},$$

och

$$\widehat{Cov(X,Y)} = \frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y}) =: \frac{S_{xy}}{n}.$$

Då blir

$$\hat{\rho} = R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

vilket kallas **Pearsons korrelationskoefficient**. Man kan räkna det

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2)(n\sum y_i^2 - (\sum y_i)^2)}}.$$

Ex.11.6.1: Man vill mäta nitratkoncentrationen av vattnet i en sjö. Man har tidigare använt ett manuelt mätningsystem men nu finns det en automatiserad metod. Om korrelationen mellan de två metoderna är stark (och positiv), börjar man använda den nya automatiserade metoden. (Data: $n = 10$, $\sum x_i = 2405$, $\sum y_i = 2503$, $\sum x_i^2 = 900.775$, $\sum y_i^2 = 919.489$, och $\sum x_i y_i = 902.475$)