F18, Multipel linjär regression.

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Multipel linjär regression

Multipel linjär regression innebär att man analyserar det linjära sambandet mellan den beroende variabeln och flera förklarande variabler. T. ex. en individs konsumtion styrs av hennes inkomst, familjeförhållanden, besparingar osv. Det kan beskrivas med hjälp av följande matematiska modell

$$kons = b_1 ink + b_2 fam + b_3 besp + \varepsilon,$$

Vid multipel linjär regression skrivs ekvationen för väntevärdet betingat på $x_1, ..., x_k$ *i populationen* generellt som

$$\mu_{y|x_1,...,x_k} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k,$$

där k är antalet förklarande variabler.

Regressionsmodellen (en enskild observation betingat på $x_1, ..., x_k$) skrivs som

$$\begin{aligned} Y &= \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \\ &= \mu_{y|x_1,...,x_k} + \varepsilon, \end{aligned}$$

där

$$\varepsilon = Y - \mu_{y|x_1,...,x_k}$$

är avvikelsen mellan Y och det betingade väntevärdet.

Det skattade observerade medelvärdet betingat på $x_1, ..., x_k$ i stickprovet skrivs som

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_2 x_k.$$

Den skattade modellen (en enskild observation betingat på $x_1, ..., x_k$) ges av

$$\begin{aligned} y &= a + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e \\ &= \hat{y} + e \end{aligned}$$

där

$$e = y - \hat{y}$$

är det observerade värdets avvikelse till medelvärdet för givna värden på $x_1, ..., x_k$.

Variansen kring medelvärdet i populationen betecknas σ_ε^2 och skattas väntevärdesriktigt med variansen kring det skattade medelvärdet i stickprovet (residualvariansen) som ges av

$$s_e^2 = MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n - (k+1)}$$

Hur skall vi skatta interceptet a och lutningskoefficienterna $b_1, b_2..., b_k$. Observera att det inte längre är en linje som skall skattas (om vi har två förklarande variabler är det ett plan, tre förklarande en kub osv).

Minsta Kvadrat Metoden

Den innebär att man vill minimera följande kvadratsumma

$$Q = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - b_1 x_1 - b_2 x_2 - ... - b_k x_k)$$

Genom partiell derivering av Q på a och $b_1, b_2, ...b_k$, respektive, får man $k+1$ ekvationer som man kan lösa ut regressionskoefficienterna från.

Man kan visa att väntevärdena för skattningarna är

$$\begin{aligned} E(a) &= \alpha \\ E(b_i) &= \beta_i, i = 1, ..., k \end{aligned}$$

och varianserna för skattningarna är

$$\begin{aligned} V(a) &= \sigma_a^2 \\ V(b_i) &= \sigma_{b_i}^2, \end{aligned}$$

som skattas i stickprovet med

$$\begin{aligned} &s_a^2 \\ &s_{b_i}^2. \end{aligned}$$

- Tolkning av $b_i, i = 1, ..., k$: Då x_i ökar en enhet ökar (minskar) y med *igenomsnitt* b_i enheter då de övriga $x-$ värdena hålls konstanta.

- Tolkning av a (interceptet): Då alla x_i är lika med noll är y *igenomsnitt* a enheter. Notera att ofta är tolkningen av interceptet a meningslös.

Exempel 1 Körner (sid 352):

$Y =$ Pris (1000 kr), $X_1 =$ Ålder (år), $X_2 =$ Körsträcka (1000 mil)

Väntevärdet *i populationen* betingat på x_1 och x_2 skrivs som

$$\mu_{y|x_1,x_2} = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

som i stickprovet skattas med

$$\hat{y} = 97.7 - 1.27 x_1 - 2.97 x_2.$$

- Tolkning av b_1 : Givet ett värde på körsträckan minskar priset med *igenomsnitt* 1270 kr då bilen åldras ett år.

- Tolkning av b_2 : Givet ett värde på åldern minskar priset med *igenomsnitt* 2970 kr då körsträckan ökar med 1000 mil.

- Tolkning av a (interceptet): Priset på en ny bil är *igenomsnitt* 97700 kr.

För en bil som är tre år äldre än en annan bil och har körts 5000 mil längre blir den genomsnittliga prisskillnaden

$$3\,(-1.27) + 5\,(-2.97) = -18.66,$$

dvs 18660 kr lägre.

Hur pass bra är punktskattningarna $y, \hat{y}, a$ och $b_i, i = 1, ..., k$. För att få en uppfattning om precisionen i skattningarna vill vi göra k.i eller hypotestest. För detta krävs modellantaganden.

- Värdena på variablerna $X_1, X_2, ..., X_k$ antas vara fixa. All slumpmässighet finns i Y.

- För varje fix kombination av $X_1, X_2, ..., X_k$ antas feltermerna ϵ_i (observationerna Y_i) vara normalfördelade med väntevärde 0 ($\mu_{y|x}$) och varians σ_ε^2.

- Feltermerna ϵ_i (observationerna Y_i) är oberoende sinsemellan.

- Homoscedastisitet. Variansen för feltermerna ϵ_i (observationerna Y_i) , σ_ε^2, är lika för varje fixed kombination av $X_1, X_2, ..., X_k$.

- Ej perfekt korrelation mellan $X-$ variablerna (multikollinearitet)

Konfidensintervall

Om feltermerna ϵ_i och observationerna Y_i är normalfördelade följer (av satsen om linjära kombinationer) att även skattningarna av regressionskoefficienterna a och $b_i, i = 1, ..., k$ måste vara normalfördelade. Alltså $a \sim N\left(\alpha; \sigma_a^2\right)$ och $b_i \sim N\left(\beta; \sigma_{b_i}^2\right)$. Ett $(1-\alpha)\,100$ % k.i. för α ges då av

$$a \pm t_{\alpha/2}\,(n-k-1)\,s_a$$

och β_i

$$b_i \pm t_{\alpha/2}\,(n-k-1)\,s_{b_i}.$$

Exempel 1 Körner (sid 352): $n = 8, s_a = 12.55, s_{b_1} = 3.64$ och $s_{b_2} = 2.11$:

Ett 95 % k.i. för α ges då av

$$\begin{aligned} & a \pm t_{0.025}\,(n-3)\,s_a \\ = \;& 97.7 \pm 2.57 \cdot 12.55 = 97.7 \pm 32.25, \end{aligned}$$

och för β_1 av

$$\begin{aligned} & b_1 \pm t_{0.025}\,(n-3)\,s_{b_1} \\ = \;& -1.27 \pm 2.57 \cdot 3.64 = -1.27 \pm 9.35, \end{aligned}$$

och för β_2 av

$$\begin{aligned} & b_2 \pm t_{0.025}\,(n-3)\,s_{b_2} \\ = \;& -2.97 \pm 2.57 \cdot 2.11 = -2.97 \pm 5.42. \end{aligned}$$

Med 95 % tillförlitlighet ligger α i intervallet $(65.4; 129.9)$, β_1 i intervallet $(-10.6; 8.1)$ och β_2 i intervallet $(-8.4; 2.4)$.

Hypotesprövning

Test av hela modellen

Vi vill undersöka om alla k oberoende variabler tillsammans *inte* kan förklara en signifikant del av variationen i Y (dvs vi har ingen regression), mot att minst en av variablerna signifikant förklarar en del av variationen i Y, dvs vi vill testa hypotesen

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

mot alternativet

$$H_1 : \text{Minst en } \beta_i \neq 0, \quad i = 1, ..., k.$$

Då feltermerna ϵ_i (observationerna Y_i) är normalfördelade följer att testvariabeln

$$\begin{aligned} F &= \frac{SSR/k}{SSE/(n-k-1)} \\ &= \frac{MSR}{MSE} \sim F(k; n-k-1) \text{ då } H_0 \text{ är sann.} \end{aligned}$$

Exempel 1 Körner (sid 352):

För test av hela modellen behöver vi $ANOVA-$tablån:

Variations-orsak	SS	f.g.	MS
Regr (R)	1843.1	$k = 2$	921.55
Res (E)	200.8	$n - k - 1 = 5$	40.16
Tot (T)	2043.9	$n - 1 = 7$	

Vi vill testa

$$H_0 : \beta_1 = \beta_2 = 0$$

mot alternativet

$$H_1 : \text{Minst en } \beta_i \neq 0, \quad i = 1, 2$$

för $\alpha = 0.01$.

Testvariabeln

$$\begin{aligned} F &= \frac{SSR/2}{SSE/(n-2-1)} \\ &= \frac{MSR}{MSE} \sim F(2; n-2-1) \text{ då } H_0 \text{ är sann.} \end{aligned}$$

H_0 förkastas om $F_{obs} > 13.3$. Det observerade testvärdet är

$$F_{obs} = \frac{SSR/2}{SSE/(8-2-1)} = \frac{1843.1/2}{200.78/5} = 22.95$$

Nollhypotesen förkastas.

Ekvivalenta tolkningar: Det kan anses statistiskt påvisat att

- det finns ett linjärt samband mellan pris och minst en av variablerna ålder och körsträcka
- variationen i ålder och/eller körsträcka förklarar en del av variationen i pris

Förklaringsgraden ges av

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{200.8}{2043.9} = 0.90$$

Tolkning: Av den totala variationen i pris förklaras 90 % av variationen i ålder och körsträcka (av modellen).

Det finns ett korrigerat R^2- värde, R^2_{adj}, som ges av

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{200.8/5}{2043.9/7} = 0.86$$

R^2_{adj} ökar nödvändigtvis inte som R^2 gör då antalet förklarande variabler ökar. Notera att man inte kan tolka R^2_{adj} som förklaringsgrad.