F19, (Multipel linjär regression forts) och F20, Chi-två test.

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Partiella t-test

Då man testar om en enskild variabel X_i skall vara med i en multipel regressionsmodel, tolkas nollhypotesen som X_i förklarar inget av variationen i Y givet att de övriga $k-1$ förklarande variablerna ingår i modellen. Mothypotesen tolkas som X_i förklarar en signifikant del av den variation i Y som återstår givet att de övriga $k-1$ förklarande variablerna ingår i modellen. Vi vill testa hypotesen

$$H_0 : \beta_i = 0, \quad i = 1, 2, ..., k$$

mot något av alternativen

$$\begin{aligned} H_1 &: \beta_i \neq 0 \\ H_1 &: \beta_i < 0 \\ H_1 &: \beta_i > 0. \end{aligned}$$

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

$$t = \frac{b_i - \beta_{i0}}{s_{b_i}} \sim t\,(n-k-1) \text{ då } H_0 \text{ är sann.}$$

Exempel 1 Körner (sid 352):

Vi vill testa

$$H_0 : \beta_1 = 0$$

mot alternativet

$$H_1 : \beta_1 \neq 0$$

för $\alpha = 0.05$. Testvariabeln

$$t = \frac{b_1 - \beta_{10}}{s_{b_1}} \sim t\,(n-2-1) \text{ då } H_0 \text{ är sann.}$$

H_0 förkastas om $|t_{obs}| > 2.57$.

Det observerade testvärdet är

$$t_{obs} = \frac{-1.27 - 0}{3.64} = -0.35$$

Nollhypotesen förkastas inte. Ekvivalenta tolkningar: Det kan inte anses statistiskt påvisat att

- ålder förklarar en del av den variation i pris som återstår efter att körsträckan fångat upp en del av variationen.
- modellen skall utökas med ålder givet att körsträcka ingår.

På motsvarande sätt kan man testa

$$H_0 : \beta_2 = 0$$

mot alternativet

$$H_1 : \beta_2 \neq 0$$

för $\alpha = 0.05$. Testvariabeln

$$t = \frac{b_2 - \beta_{20}}{s_{b_2}} \sim t\,(n-2-1) \text{ då } H_0 \text{ är sann.}$$

H_0 förkastas om $|t_{obs}| > 2.57$. Det observerade testvärdet är

$$t_{obs} = \frac{-2.97 - 0}{2.11} = -1.41$$

Nollhypotesen förkastas inte. Det kan inte anses statistiskt påvisat att körsträckan förklarar en del av den variation i pris som återstår efter att ålder fångat upp en del av variationen.

Multikollinearitet

R^2 ökar alltid då antalet förklarande variabler ökar. Samtidigt minskar ofta säkerheten i skattningarna av enskilda regressionskoefficienter. Detta gäller särskilt om man har hög korrelation mellan de förklarande variablerna (multikollinearitet).

Modellval

Antag att vi vill bygga en modell och har 25-30 förklarande variabler att välja bland. Skall vi ta med samtliga variabler i modellen? Om inte, vilka skall vi välja att ingå i vår modell? Det beror lite på vad vi vill ha modellen till.

- Vill vi ha den till att skatta enskilda förklarande variablers effekt på $y-$ variabeln. Ta med *väsentligt* förklarande variabler i modellen. Har de förklarande variablerna hög kausalitet med underökningsvariabeln? Intressant diskussion (AJÅ sid 92-95).

- Vill vi enbart ha den till att göra prognoser med. Kausalitet och krav på *signifikant* förklarande variabler är inte lika viktigt. Det viktiga är att skatta en modell som ger bra prognoser.

- Man skall inte ösa på med förklarande variabler i modellen enbart för att R^2 ökar. Vi får fler parametrar att skatta och osäkerheten i skattningarna ökar.

- *Simplicitetsprincipen*. Försök att skatta en modell med så få förklarande variabler som möjligt. Oftast mindre osäkerhet i parameterskattningarna och oftast bättre prognoser.

Några (statistiska) kriterier att följa vid val av modell:

- Så stort R^2_{adj} som möjligt (liten residualspridning s_e)

- Undersök om modellen håller som helhet (F-test)

- Ta enbart med variabler som är signifikant förklarande, dvs $\beta_i \neq 0$ (t-test).

- *Simplicitetsprincipen*. Även om en variabel är signifikant förklarande men R^2_{adj} (R^2) ökar marginellt, kanske man skall överväga att utelämna den varibeln.

- Gör en residualanalys för att se om modellen håller.

Använd t ex menyn "Best subset regression" i MINITAB.

Prognoser (prediktioner)

På liknande sätt som i enkel linjär regression kan man göra prognoser i form av punktskattningar och intervallskattningar dels för "linjen" (medelvärdet) och dels för enskilda observationer. Punktskattningen ges av

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

både för medelvärdet $E(Y|x_1, x_2, ..., x_k) = \mu_{y|x_1,...,x_k}$ och en enskild observation $Y|x_1, x_2, ..., x_k$.

Den skattade standardavvikelsen för punktskattningen med avseende på $E(Y|x_1, x_2, ..., x_k)$ skrivs som

$$s_{\hat{y}|x_1,...,x_k},$$

och den skattade standardavvikelsen för punktskattningen med avseende på $Y|x_1, x_2, ..., x_k$ skrivs som

$$s_{y|x_1,...,x_k} = sqrt\left(s^2_{\hat{y}|x_1,...,x_k} + s^2_e\right).$$

Exempel 1 Körner (sid 352):

Ge punktskattning och konfidensintervall för det genomsnittliga priset och prognosintervall för priset för en enskild bil då $x_1 = 10$ år och $x_2 = 16$ tusen mil. Punktskattningen blir

$$\begin{aligned} \hat{y} &= a + b_1 x_1 + b_2 x_2 \\ &= 97.73 - 1.27 \cdot 10 - 2.97 \cdot 16 = 37.5, \end{aligned}$$

och de skattade standardavvikelserna enl. MINITAB-utskrift (Körner sid 358)

$$\begin{aligned} s_{\hat{y}|x_1,...,x_k} &= 4.91 \\ s_{y|x_1,...,x_k} &= sqrt\left(4.91^2 + 6.337^2\right) = 8.02 \end{aligned}$$

Ett 95 %-igt k.i. för $E(Y|x_1 = 10, x_2 = 16)$ ges av

$$\begin{aligned} & \hat{y} \pm t_{0.025}(n-k-1)\, s_{\hat{y}|x_1,x_2} \\ = \; & 37.5 \pm 2.57 \cdot 4.91 = 37.5 \pm 12.6, \end{aligned}$$

och ett 95 %-igt k.i. för $Y|x_1 = 10, x_2 = 16$ ges av

$$\begin{aligned} & \hat{y} \pm t_{0.025}(n-k-1)\, s_{y|x_1,x_2} \\ = \; & 37.5 \pm 2.57 \cdot 8.02 = 37.5 \pm 20.6, \end{aligned}$$

Med 95 % tillförlitlighet ligger $E(Y|x_1 = 10, x_2 = 16)$ i intervallet $(24.9; 50.1)$ och $Y|x_1 = 10, x_2 = 16$ i intervallet $(16.9; 58.1)$.

Dummy-variabel (Indikator-variabel)

Hittils har vi använt kvantitativa variabler i regressionsmodellen. Det kan även vara av intresse att ha med kvalitativa (går ej att rangordna) förklarande variabler, t.ex kön, säsong, nationalitet, yrkeskategori, m.m. Vi använder oss då av en dummyvariabel som är dikotom och antar värdet 0 eller 1. Den kvalitativa variabeln kan anta hur många värden som helst, men här begränsar vi oss till fallet där den enbart antar två värden.

Exempel 1 Körner (sid 352):

$Y =$ Pris (1000 kr), $X_1 =$ Ålder (år), $X_2 =$ Säljare (med värdena "privatperson" $= 0$ och "bilfirma" $= 1$). Den skattade regressionekvationen ges fortfarande av

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

som blir

$$\hat{y} = a + b_1 x_1$$

om bilen säljs av en privatperson och

$$\hat{y} = (a + b_2) + b_1 x_1$$

om bilen säljs av en bilfirma. Regressionskoefficienten b_2 tolkas alltså som genomsnittliga skillnaden i pris då bilen säljs privat och då bilen säljs av en bilfirma, givet att bilarna är lika gamla.

Exempel 2 Körner (sid 364):

Den skattade modellen blir

$$\hat{y} = 97.2 - 5.79 x_1 + 6.56 x_2.$$

Tolkning b_2 : En bil är 6560 kr dyrare igenomsnitt om den säljs av en bilfirma, givet samma årsmodell på bilarna.

χ^2- test

Olika typer av test som ibland kallas fördelningsfria test då fördelningen för testvariabeln inte beror på fördelningen för undersökningsvariablerna.

Fördelningstest

Vi vill undersöka om frekvenserna följer en viss teoretisk sannolikhetsfördelning.

Exempel:

Ett klockföretag skall lansera en ny klocka och vill undersöka om presumtiva konsumenter har specifika preferenser vad gäller färg på armbandet, eller är alla fyra färger lika populära. Dvs följer antalet sålda klockor samma fördelning med avseende på armbandsfärg? Vi vill testa följande hypoteser:

H_0 : $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ (Konsumenterna väljer slumpm. klocka)

H_1 : Alla övriga alternativ på p_1, p_2, p_3, p_4 (Konsumente väljer ej slumpm. klocka),

där p_1, p_2, p_3 och p_4 är sannolikheten att konsumenten väljer klocka med färg ett, två, tre eller fyra, respektive.

Vi har samlat följande data från 80 slumpm. valda presumtiva konsumenter:

Färg	Obs antal sålda klockor	Förv antal sålda klock (under H_0)
1	12	$80 \cdot \frac{1}{4} = 20$
2	40	$80 \cdot \frac{1}{4} = 20$
3	8	$80 \cdot \frac{1}{4} = 20$
4	20	$80 \cdot \frac{1}{4} = 20$

I testvariabeln jämförs de observerade frekvenserna med de förväntade frekvenserna under nollhypotesen på följande sätt:

$\chi^2 = \sum \frac{(O-E)^2}{E} \sim approx\ \chi^2(a-1) -$ fördelad då H_0 är sann,

där a är antalet kategorier, O är antalet observerade frekvenser och E är antalet förväntade frekvenser under nollhypotesen. Om skillnaden mellan de observerade frekvenserna och de förväntade frekvenserna blir för stora förkastas nollhypotesen.

De förväntade frekvenserna får inte vara alltför små. Följande tumregel gäller:

- Ingen förväntad frekvens får understiga 1.
- Högst 20 % av de förväntade frekvenserna får understiga 5.

Signifikansnivån $\alpha = 0.05$, dvs H_0 förkastas om $\chi^2_{obs} > 7.81$. Det observerade testvärdet är

$$\chi^2_{obs} = \frac{(12-20)^2}{20} + \frac{(40-20)^2}{20} + \frac{(8-20)^2}{20} + \frac{(20-20)^2}{20} = 3$$

Då $30.4 > 7.81$ förkastas nollhypotesen. Det kan anses statistiskt påvisat att det föreligger skillnad i andel sålda klockor (i preferens av armbandsfärg).

Uppgift 902 (Körner):

Man vill undersöka fördelningen för fyra möjliga utfall av två egenskaper. Enligt Mendels lagar är fördelningen för de fyra utfallen 9:3:3:1. I ett genetiskt experiment erhöll man följande data:

Utfall	Frekvenser
U_1	82
U_2	36
U_3	30
U_4	12

Ger data stöd åt Mendels lagar. Låt $p_i, i = 1, 2, 3, 4$, vara sannolikheten för U_i.

Vi vill då testa följande hypoteser:

H_0 : $p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$
(Fördeln. för utfallen följer Mendels lagar)

H_1 : Alla övriga alternativ på p_1, p_2, p_3, p_4
(Fördeln. för utfallen följer ej Mendels lagar),

för signifikansnivån $\alpha = 0.05$.

Utfall	Obs antal	Förv antal (under H_0)
U_1	82	$160\frac{9}{16} = 90$
U_2	36	$160\frac{3}{16} = 30$
U_3	30	$160\frac{3}{16} = 30$
U_4	12	$160\frac{1}{16} = 10$
Summa	160	160

Testvariabeln är

$$\chi^2 = \sum \frac{(O-E)^2}{E} \sim approx\ \chi^2(a-1) -$$ fördelad då H_0 är sann.

H_0 förkastas om $\chi^2_{obs} > 7.81$. Det observerade testvärdet är

$$\chi^2_{obs} = \frac{(82-90)^2}{90} + \frac{(36-30)^2}{30} + \frac{(30-30)^2}{30} + \frac{(12-10)^2}{10}$$
$$= 2.31.$$

Då $2.31 < 7.81$ förkastas ej nollhypotesen. Det kan ej anses statistiskt påvisat att fördelningen för utfallen av de två egenskaperna ej följer Mendels lagar.