STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

9 december 2022 14–19

Examinator: Ola Hössjer, tel. 070/672 12 18, ola@math.su.se

Uppgift 1

a) För enkel linjär regression testas en enda effektparameter β, så antalet frihetsgrader för Regression är 1. Eftersom modellen innehåller totalt 2 parametrar är antalet frihetsgrader för Residual $N-2=5$. Det ger medelkvadratsummor

$$\begin{aligned} \text{Mkvs(Regression)} &= \text{Kvs(Regression)} = 9.2, \\ \text{Mkvs(Residual)} &= \text{Kvs(Residual)}/5 = 5.75/5 = 1.15, \end{aligned}$$

och en

$$\text{F-kvot} = \frac{9.2}{1.15} = 8$$

som överstiger $F_{0.05}(1,5) = 6.6$. Vi kan alltså förkasta H_0 på nivån 5% och konstatera att trädstorlek har en signifikant inverkan på graden av svampangrepp.

b) Grund- och hypotesmodellernas minsta kvadrat-skattningar av $\mu_i = \tilde{\alpha} + \beta(x_i - \bar{x})$ är

$$\begin{aligned} \hat{\mu}_i &= \hat{\tilde{\alpha}} + \hat{\beta}(x_i - \bar{x}) = \bar{Y} + \hat{\beta}(x_i - \bar{x}), \\ \hat{\hat{\mu}}_i &= \hat{\tilde{\alpha}} = \bar{Y}. \end{aligned}$$

Det ger

$$\begin{aligned} \text{Kvs(Regression)} &= \textstyle\sum_{i=1}^7 (\hat{\mu}_i - \hat{\hat{\mu}}_i)^2 \\ &= \textstyle\sum_{i=1}^7 \left[\hat{\beta}(x_i - \bar{x})\right]^2 \\ &= \hat{\beta}^2 \textstyle\sum_{i=1}^7 (x_i - \bar{x})^2 \\ &= \hat{\beta}^2 \left[(-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2\right] \\ &= 28\hat{\beta}^2. \end{aligned}$$

Med extrainformationen $\hat{\beta} < 0$ fås slutligen

$$\hat{\beta} = -\sqrt{\frac{\text{Kvs(Regression)}}{28}} = -\sqrt{\frac{9.2}{28}} = -0.573.$$

Uppgift 2

a) Vi börjar med att fylla i antalet frihetsgrader f i den fullständiga modellens variansanalystabell, med förkortningarna K1, K2 och K3 för de tre katalysatorerna. Eftersom varje katalysator bidrar med en regressionsparameter så har motsvarande variationskälla en frihetsgrad. Det ger följande tabell:

Variationskälla	f	Kvs
K1	1	4.1
K2	1	6.2
K3	1	2.1
Residual	6	2.8
Totalt	9	15.2

b) I första FS-steget testas tre olika grundmodeller, var och en med en förklarande variabel, med ett F-test mot en hypotesmodell som bara innehåller intercept. Eftersom K2 har störst kvadratsumma börjar vi med att undersöka F-kvoten för delmodellen som endast har K2 som förklarande variabel. Om vi använder denna delmodell som grundmodell får vi enligt ledningen

$$\begin{aligned}
\text{Kvs(Regression)} &= \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 \\
&= \text{Kvs(K2)} \\
&= 6.2, \\
\text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\
&= \text{Kvs(K1)} + \text{Kvs(K3)} + \text{Kvs(Residual)}_{\text{fullst}} \\
&= \text{Kvs(Total)} - \text{Kvs(K2)} \\
&= 15.2 - 6.2 \\
&= 9.0,
\end{aligned}$$

där $\boldsymbol{Y}$ är observationsvektorn, och $\hat{\boldsymbol{\mu}}$, $\hat{\hat{\boldsymbol{\mu}}}$ skattningar av dess väntevärde enligt grund- respektive hypotesmodellen. Eftersom antalet frihetsgrader för Regression och Residual är 1 respektive $1 + 1 + 6 = 8$ får vi en

$$\begin{aligned}
\text{F-kvot} &= \frac{\text{Kvs(Regression)}/1}{\text{Kvs(Residual)}/8} \\
&= \frac{6.2}{9.0/8} \\
&= 5.51
\end{aligned}$$

som överstiger $F_{0.05}(1, 8) = 5.3$. Därför ger K2 ett signifikant bidrag i första steget av FS-schemat.

De andra två delmodellerna med bara K1 respektive K3 som förklarande variabler har också 1 frihetsgrad för Regression och 8 frihetsgrader för Residual. Eftersom deras Kvs(Regression) = Kvs(K1) respektive Kvs(Regression) = Kvs(K3) är lägre och deras Kvs(Residual) = Kvs(Total) − Kvs(Regression)

är högre jämfört med modellen som bara har K2 som förklarande variabel, så kommer F-kvoterna för båda dessa delmodeller att understiga 5.51. Därför väljer vi inte heller någon av dessa båda delmodeller, utan delmodell K1 väljs som förklarande variabel i FS-schemats första steg.

b) Eftersom K2 vales i a) så stannar FS-schemat inte efter detta första steg. Istället går vi vidare till ett andra steg där hypotesmodellen med bara K2 som förklarande variabel testas mot två olika grundmodeller, som även inkluderar K1 eller K3. Eftersom K1 har större kvadratsumma än K3, och alla prediktorerna är ortogonala, räcker det att i andra steget undersöka om K1 ska tas med utöver K2, av samma skäl som att det i a) räckte att testa grundmodellen med K2 mot hypotesmodellen med endast intercept. Med K1 och K2 i grundmodellen och bara K2 i hypotesmodellen får vi då, på grund av ortogonaliteten mellan prediktorerna (se ledningen), att

$$\begin{aligned} \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 &= \text{Kvs(Regression)}_{\text{K1+K2}} - \text{Kvs(Regression)}_{\text{K2}} \\ &= [\text{Kvs(K1)} + \text{Kvs(K2)}] - \text{Kvs(K2)} \\ &= \text{Kvs(K1)} \\ &= 4.1, \\ \text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(K3)} + \text{Kvs(Residual)}_{\text{fullst}} \\ &= 2.1 + 2.8 \\ &= 4.9, \end{aligned} \tag{1}$$

och en

$$\begin{aligned} \text{F-kvot} &= \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/1}{\text{Kvs(Residual)}/(1+6)} \\ &= \frac{4.1}{4.9/7} \\ &= 5.86, \end{aligned} \tag{2}$$

som överstiger $F_{0.05}(1, 7) = 5.6$. Således väljs en modell med både K1 och K2 som förklarande variabler som modell i FS-schemats andra steg.

Uppgift 3

a) Låt

$$\bar{Y}_{\cdot\cdot} = \frac{\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot} + \bar{Y}_{3\cdot}}{3} = \frac{38.1 + 38.4 + 37.8}{3} = 38.1$$

vara det totala stickprovsmedelvärdet för hela studien. Beteckna de två variationskällorna

$$\begin{aligned} \text{MV} &= \text{Mellan vaccin}, \\ \text{IV} &= \text{Inom vaccin}, \end{aligned}$$

med antal frihetsgrader $3-1=2$ respektive $3(6-1)=15$. De två medelkvadratsummorna ges av

$$\begin{aligned}\text{Mkvs(MV)} &= \tfrac{1}{3-1}\sum_{i=1}^{3}\sum_{j=1}^{6}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2\\ &= \tfrac{6}{3-1}\left[(\bar{Y}_{1\cdot}-\bar{Y}_{\cdot\cdot})^2+(\bar{Y}_{2\cdot}-\bar{Y}_{\cdot\cdot})^2+(\bar{Y}_{3\cdot}-\bar{Y}_{\cdot\cdot})^2\right]\\ &= 3\left[0^2+0.3^2+(-0.3)^2\right]\\ &= 0.54\end{aligned}$$

och

$$\begin{aligned}\text{Mkvs(IV)} &= \tfrac{1}{3(6-1)}\sum_{i=1}^{3}\sum_{j=1}^{6}(Y_{ij}-\bar{Y}_{i\cdot})^2\\ &= \tfrac{5}{15}(s_1^2+s_2^3+s_3^3)\\ &= \tfrac{1}{3}(0.25^2+0.31^2+0.19^2)\\ &= 0.0649.\end{aligned}$$

Det ger en

$$\text{F-kvot} = \frac{0.54}{0.0649} = 8.32$$

som överstiger $F_{0.05}(2,15)=3.7$. Vi kan alltså förkasta nollhypotesen att de tre vaccinerna ger samma grad av biverkning.

b) Låt $\Delta=(\mu_1+\mu_2)/2-\mu_3$ vara den förväntade biverkningsminskningen om ett vaccinbyte görs, med punktskattning

$$\hat{\Delta}=\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)-\hat{\mu}_3=\frac{1}{2}(\bar{Y}_{1\cdot}+\bar{Y}_{2\cdot})-\bar{Y}_{3\cdot}=\frac{1}{2}(38.1+38.4)-37.8=0.45.$$

c) Om $\sigma^2=\text{Var}(Y_{ij})$ är feltermsvariansen, följer att

$$\text{Var}(\bar{Y}_{i\cdot})=\frac{\sigma^2}{6}$$

och

$$\text{Var}(\hat{\Delta})=\frac{1}{2^2}\text{Var}(\bar{Y}_{1\cdot})+\frac{1}{2^2}\text{Var}(\bar{Y}_{2\cdot})+\text{Var}(\bar{Y}_{3\cdot})=\frac{\sigma^2}{6}(\frac{1}{4}+\frac{1}{4}+1)=\frac{\sigma^2}{4}.$$

Det ger ett medelfel

$$d=\sqrt{\widehat{\text{Var}}(\hat{\Delta})}=\frac{\hat{\sigma}}{2}=\frac{\sqrt{\text{Mkvs(IV)}}}{2}=\frac{\sqrt{0.0649}}{2}=0.1274.$$

Uppgift 4

a) Vi utnyttjar ledningen för att ge ett uttryck för kovariansfunktionen γ_k då $k=1,2$. Vi börjar med $k=1$ och ser att

$$\begin{aligned}\gamma_1 &= \text{Cov}(X_t,X_{t+1})=\text{Cov}(X_t,\phi_1X_t+\phi_2X_{t-1}+\varepsilon_{t+1})\\ &= \phi_1\text{Cov}(X_t,X_t)+\phi_2\text{Cov}(X_t,X_{t-1})+\text{Cov}(X_t,\varepsilon_{t+1}) \qquad (3)\\ &= \phi_1\gamma_0+\phi_2\gamma_1.\end{aligned}$$

Genom att lösa (3) med avseeende på γ_1 får vi

$$\gamma_1 = \phi_1\gamma_0/(1-\phi_2) \Longleftrightarrow \rho_1 = \gamma_1/\gamma_0 = \phi_1/(1-\phi_2). \tag{4}$$

Vi fortsätter med $k = 2$ och noterar att

$$\begin{aligned} \gamma_2 &= \mathrm{Cov}(X_t, X_{t+2}) = \mathrm{Cov}(X_t, \phi_1 X_{t+1} + \phi_2 X_t + \varepsilon_{t+2}) \\ &= \phi_1\gamma_1 + \phi_2\gamma_0 = \gamma_0(\phi_1\rho_1 + \phi_2) \\ &= \gamma_0(\phi_1^2/(1-\phi_2) + \phi_2), \end{aligned} \tag{5}$$

där vi i sista ledet utnyttjade (4). Division med γ_0 i båda leden av (5) leder slutligen till

$$\rho_2 = \gamma_2/\gamma_0 = \phi_1^2/(1-\phi_2) + \phi_2.$$

b) Låt $\bar{y} = \hat{\mu} = \sum_{t=1}^{T} y_t/T$ vara stickprovsmedelvärdet av den observerade tidsserien. Skattningen av kovariansfunktionen ges av

$$\hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y}),$$

vilket i sin tur ger en skattning

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

av autokorrelationsfunktionen (korrelogrammet).

c) Eftersom $Y_t = \mu + \varepsilon_t$ är oberoende under nollhypotesen så gäller approximativt, för stora T, att $\hat{\rho}_1$ och $\hat{\rho}_2$ är oberoende med $\sqrt{T}\hat{\rho}_k \sim N(0,1)$. Det medför att vi approximativt har

$$Q = T(\hat{\rho}_1^2 + \hat{\rho}_2^2) \sim \chi^2(2)$$

under H_0. Ett hypotestest som förkastar H_0 då $Q > \chi^2_{0.05}(2) = 5.99$ har alltså approximativt signifikansnivån 5%.

Uppgift 5

a) Enligt definitionen av den multipla linjära regressionsmodellen har responsvariablerna väntevärden

$$\mu_i = \tilde{\alpha} + \sum_{j=1}^{m}\beta_j(x_{ji} - \bar{x}_j), \quad i = 1, \ldots, N,$$

där x_{ji} är förklarande variabel nummer j för observation i, $\bar{x}_j = \sum_{i=1}^{N} x_{ji}/N$, $\tilde{\alpha}$ är (det centrerade) interceptet och $\beta_1, \ldots, \beta_m$ effektparametrar.

b) Förklaringsgraden R^2 definieras som den andel av den *totala* variationen som härrör från de förklarande variablerna (*regressions*delen). Det uttrycks med kvadratsummor som

$$R^2 = \frac{\text{Kvs(Regression)}}{\text{Kvs(Total)}} = \frac{\sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}. \tag{6}$$

c) Från definitionen på korrelationskoefficienten och ledningen i uppgiften följer att

$$\begin{aligned} \text{Corr}(\boldsymbol{Y}, \hat{\boldsymbol{\mu}})^2 &= \frac{\left[\frac{1}{N}\sum_{i=1}^N (Y_i - \bar{Y})(\hat{\mu}_i - \bar{Y})\right]^2}{\frac{1}{N}\sum_{i=1}^N (Y_i - \bar{Y})^2 \cdot \frac{1}{N}\sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2} \\ &= \frac{\left[\sum_{i=1}^N (Y_i - \bar{Y})(\hat{\mu}_i - \bar{Y})\right]^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2 \cdot \sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2}. \end{aligned} \tag{7}$$

Genom uppdelningen $Y_i - \bar{Y} = (Y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{Y})$ ser vi att summan i täljaren i (7) kan förenklas till

$$\begin{aligned} \sum_{i=1}^N (Y_i - \bar{Y})(\hat{\mu}_i - \bar{Y}) &= \sum_{i=1}^N (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{Y}) + \sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2 \\ &= \sum_{i=1}^N (Y_i - \hat{\mu}_i)\hat{\mu}_i + \sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2 \\ &= \sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2, \end{aligned} \tag{8}$$

där vi i andra steget utnyttjade att $\sum_i Y_i = \sum_i \hat{\mu}_i = N\bar{Y}$, och i sista steget att $\boldsymbol{Y} - \hat{\boldsymbol{\mu}}$ är ortogonal mot $\hat{\boldsymbol{\mu}}$. Det följer av att $\hat{\boldsymbol{\mu}}$ är den ortogonala projektionen av $\boldsymbol{Y}$ ned på det underrum av $\mathbb{R}^N$ som spänns upp av de $m+1$ kolumnerna i designmatrisen $\boldsymbol{A}$. Genom att sätta in (8) i (7) ser man att $\text{Corr}(\boldsymbol{Y}, \hat{\boldsymbol{\mu}})^2$ är identisk med R^2 i (6), eftersom en faktor $\sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2$ förkortas bort i täljare och nämnare.