

Bayesiansk statistik, 732g43, 7.5 hp

Moment 2 - Linjär regressionsanalys

Bertil Wegmann

STIMA, IDA, Linköpings universitet

- Bayesiansk linjär regression utan förklaringsvariabler
- Bayesiansk enkel linjär regression
- Bayesiansk multipel linjär regression
- Kod_Moment2.R (kan laddas ned på Lisam)

- Modell:

$$Y_1, \dots, Y_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$$
$$\mu = \beta_0$$

där både medelvärdet μ och variansen σ^2 är okända.

- Modellen kan skrivas som en linjär regressionsmodell utan förklaringsvariabler:

$$y_i = \beta_0 + \epsilon_i$$
$$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

- Prior:

$$p(\mu, \sigma)$$

- Om μ och σ antas oberoende apriori (boken):

$$p(\mu, \sigma) = p(\mu) p(\sigma).$$

- Priorn för μ och σ^2 kan också specificeras enligt:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2) p(\sigma^2)$$

- Standard icke-informativ prior är en uniform prior för $(\mu, \ln \sigma)$:

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- Fördelen med standard icke-informativ prior:

- betingade posteriorn $p(\mu|\sigma^2)$ och den marginella posteriorn $p(\sigma^2)$ följer kända fördelningar.
- ger acceptabla resultat om man har mycket data.

- Nackdelar:

- om man har lite data bör man specificera en rimligare prior (se oberoende priors i boken), eftersom prior blir mer viktig vid lite data.

- Modell för linjär regression:

$$Y_1, \dots, Y_n | \mu, \sigma^2, \mathbf{x} \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

$$\mu = \boldsymbol{\beta} \mathbf{x}',$$

där variansen σ^2 är okänd och vektor med förklaringsvariabler $\mathbf{x} = (1 \ x_1 \ \dots \ x_k)$ och parametrar $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_k)$ i stället för endast $\mu = \beta_0$ i föregående modell.

- Om $\boldsymbol{\beta}$ och σ antas oberoende apriori (boken):

$$p(\boldsymbol{\beta}, \sigma) = \prod_{j=0}^k p(\beta_j) p(\sigma).$$

- Standard icke-informativ prior är här en uniform prior för parametrarna $(\boldsymbol{\beta}, \ln \sigma)$:

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{x}) \propto \frac{1}{\sigma^2}.$$

- Fördelen med den standard icke-informativa priorn $p(\beta, \ln \sigma)$:
 - betingade posteriorn $p(\beta|\sigma^2)$ och den marginella posteriorn $p(\sigma^2)$ följer även här kända fördelningar.
 - ger acceptabla resultat om man har mycket data jämfört med antalet förklaringsvariabler i x .
- Nackdelar:
 - om man har lite data eller många förklaringsvariabler, så bör man specificera en rimligare prior (se oberoende priors i boken).

- Ett slumpmässigt urval av 32 bilar har dragits från *1974 Motor Trend US magazine*, se datamaterialet *mtcars* från R:s dataexempel.
- Slutgiltigt mål: multipel linjär regressionsanalys med den beroende variabeln $y = \text{miles}/(\text{US}) \text{ gallon}$ för en bils bränsleförbrukning.
- 1 *miles/gallon* motsvarar ungefär 23,52 liter per mil. Transformera om y till liter per mil.
- Modell utan förklaringsvariabler:

$$Y_1, \dots, Y_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

$$\mu = \beta_0$$

där både medelvärdet μ och variansen σ^2 är okända.

- Oberoende priors för μ och σ (använd t.ex. webbverktyget för att elicitera priorn):

$$\mu \sim N(1, 1^2)$$

$$\sigma \sim \text{Uniform}(0, 2)$$

- Plotta priors för μ och σ för att plotta dina antaganden apriori.
- Plotta data y för att se vad dina priors ger för apriori information om antal liter per mil.
 - 1 Dra många värden på μ och σ från priorfördelningarna.
 - 2 Dra betingade värden $y|\mu, \sigma^2$ givet värdena i punkt 1.
- Verkar dina priors rimliga? Om inte, ändra priors för μ och σ tills du blir nöjd.
- Kontakta experter, t.ex. bilhandlare, bilverkstäder, bilföreningar, etc., om du behöver hjälp med att elicitera en rimlig prior.

- Kvadratisk approximation med funktionen **map** i R-paketet **rethinking**, #R kod 2.2
- Problem med kvadratisk approximation för σ , eftersom standardavvikelse- eller variansparametrar har en tendens att vara skeva åt höger.
- Kvadratisk approximation blir i bland bättre för $\ln \sigma$. #R kod 2.3
- Marginell posteriorfördelning för σ bestäms genom att antilogaritmera posteriorfördelningen för $\ln \sigma$:

$$\sigma_i = \exp [(\ln \sigma)_i],$$

där i är den i :te samplade dragningen från respektive posteriorfördelning.

- Jämför posteriorfördelningarna för σ med respektive kvadratisk approximation för σ och $\ln \sigma$.
- Om $\ln \sigma | y_1, \dots, y_n \sim N(\mu_n, \sigma_n)$, så följer posteriorfördelningen för σ en *log-normal* fördelning med parametrar μ_n och σ_n .

- Posteriorn för \hat{y} i linjär regression utan förklaringsvariabler är posteriorn för medelvärdet μ .
- Posterior prediktiv fördelning för nya observationer \tilde{y} givet data y_1, \dots, y_n , $p(\tilde{y}|y)$. **#R kod 2.4**
- Modellutvärdering med replikerade data (*in-sample fit*):
 - plotta $p(\tilde{y}|y)$ genom att dra värden från posteriorfördelningen $(\mu, \sigma) | y_1, \dots, y_n$:
 - 1 Dra många värden på (μ, σ) från posteriorfördelningen.
 - 2 Dra nya observationer \tilde{y} från $\tilde{y} | \mu, \sigma^2 \sim N(\mu, \sigma^2)$ givet (μ, σ) i punkt 1.

- Även om μ och σ antas vara oberoende apriori, så tillåts dom vara beroende aposteriori.
- Funktion för kovariansen mellan μ och σ från posteriorn: `vcov()`.
- Korrelationsmatris: `cov2cor()`.
- Dra posteriorvärden för μ och σ direkt från multivariat normalfördelning (kvadratisk approximation):

`mvrnorm(n = Nsamples, mu = coef(), Sigma = vcov())`

■ Modell:

$$Y_1, \dots, Y_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

■ Prior:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

$$p(\mu | \sigma^2) \propto c$$

$$p(\sigma^2) \propto (\sigma^2)^{-1}$$

■ Betingad posterior för $\mu | \sigma^2$:

$$p(\mu | \sigma^2, y_1, \dots, y_n) \propto p(y_1, \dots, y_n | \mu, \sigma^2) p(\mu | \sigma^2)$$

$$\mu | \sigma^2, y_1, \dots, y_n \sim N(\bar{y}, \sigma^2 / n)$$

- Marginell posterior för σ^2 :

$$p(\sigma^2 | y_1, \dots, y_n) = \int p(\mu, \sigma^2 | y_1, \dots, y_n) d\mu$$

$$= \int p(y_1, \dots, y_n | \mu, \sigma^2) p(\mu, \sigma^2) d\mu$$

$$\implies \sigma^2 | y_1, \dots, y_n \sim \text{Inv} - \chi^2(n-1, s^2),$$

där s^2 är urvalsvariansen för data y_1, \dots, y_n .

- $\text{Inv} - \chi^2(n-1, s^2)$ är en skalad invers χ^2 -fördelning:

$$\frac{(n-1)s^2}{\sigma^2} | y_1, \dots, y_n \sim \chi^2(n-1).$$

■ Dra värden från posteriorfördelningen för σ^2 :

- 1 Dra ett värde $f(\sigma^2) = \frac{(n-1)s^2}{\sigma^2}$ från $\chi^2(n-1)$. (använd funktion `rchisq` i R)
- 2 Beräkna det dragna värdet för σ^2 , givet värdet $f(\sigma^2)$ i punkt 1, enligt:

$$\sigma^2 = \frac{(n-1)s^2}{f(\sigma^2)}$$

- 3 Upprepa denna procedur många gånger för att få många dragna värden från posteriorfördelningen för σ^2 .

■ Dra värden från den betingade posteriorfördelningen $\mu|\sigma^2$:

$$\mu_j|\sigma_j^2, y_1, \dots, y_n \sim N(\bar{y}, \sigma_j^2/n),$$

där j är den j :te samplade dragningen från respektive posteriorfördelning för σ^2 och $\mu|\sigma^2$.

- Enkel linjär regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

- Modellen kan skrivas som:

$$Y_1, \dots, Y_n | \mu_i, \sigma^2 \stackrel{iid}{\sim} N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_i,$$

där parametrarna β_0 , β_1 och variansen σ^2 är okända.

- Prior:

$$p(\beta_0, \beta_1, \sigma)$$

- Om $(\beta_0, \beta_1, \sigma)$ antas oberoende apriori (boken):

$$p(\beta_0, \beta_1, \sigma) = p(\beta_0) p(\beta_1) p(\sigma).$$

- Enkel linjär regressionsanalys med den beroende variabeln $y = \text{liter} / \text{mil}$ för en bils bränsleförbrukning och förklaringsvariabeln $x = \text{vikt}$ i ton för en bil.
- 1 *miles/gallon* motsvarar ungefär 23,52 liter per mil. Transformera om y till liter per mil.
- 1 *pound* motsvarar ungefär 0,45 kilo. Transformera om x till ton.
- Modell:

$$Y_1, \dots, Y_n | \mu_i, \sigma^2 \stackrel{iid}{\sim} N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_i$$

- Oberoende priors för $(\beta_0, \beta_1, \sigma)$:

$$\beta_0 \sim N(1, 1^2)$$

$$\beta_1 \sim N(0.5, 1^2)$$

$$\ln \sigma \sim N(0, 0.5^2)$$

- Kvadratisk approximation med funktionen **map** i R-paketet **rethinking**. **#R kod 2.6**
- Sammanfattning av posteriorn sker oftast genom att presentera tabeller och plottar över posteriorresultatet.
- Plottar av posteriorn ger oftast mer information om posteriorn än vad tabeller ger. All osäkerhet om olika kvantiteter i modellen kan plottas men inte återges i tabeller.
- Man kan ge mer viktning till tabeller när man blir mer van vid att tolka posteriorresultatet.
- Typisk tabell inkluderar posterior medelvärde, standardavvikelsen och kredibilitetsintervall (t.ex. 90.9 % och 95.2 %).
- Oddset för positiv eller negativ lutning kan beräknas för lutningsparametern β_1 . **#R kod 2.7**

- Hög korrelation mellan interceptet β_0 och lutningen β_1 ($r_{\beta_0, \beta_1} = -0.957$) kan medföra svårigheter att skatta modellen i mer komplicerade modeller med fler förklaringsvariabler \Rightarrow Centrera eller standardisera förklaringsvariablerna. **#R kod 2.8**
- Om inte möjliga värden skiljer sig avsevärt mellan förklaringsvariablerna, så räcker det med centrerings $x_c = x - \bar{x}$. Annars är det bättre med standardisering $x_s = \frac{x - \bar{x}}{s_x}$.
- Skattade modeller med standardiserade förklaringsvariabler kan vara mer svårtolkade, men man kan konvertera tillbaka till estimationsresultat på originalskala för förklaringsvariablerna.
- Fördelen med standardiserade förklaringsvariabler är att man kan jämföra magnituderna på lutningsparametrarna för förklaringsvariablerna, eftersom förklaringsvariablernas värden är standardiserade till samma skala. Viktigt om mycket blir signifikant pga mycket data.

- Plotta den skattade regressionslinjen från map (maximum a posteriori) skattningarna.
- Alla regressionslinjer från alla posteriordragningar för β_0 och β_1 kan plottas enligt:

$$\mu_{ij} = \beta_{0j} + \beta_{1j}x_i,$$

där i gäller för observation i och j är den j :te samplade dragningen från posteriorfördelningen för (β_0, β_1) .

- Posteriorfördelningen för förväntad bensinförbrukning för en bil med vikt 1.5 ton:

$$\mu_{1.5,j} = \beta_{0j} + \beta_{1j} \cdot 1.5,$$

där j är den j :te samplade dragningen från posteriorfördelningen för (β_0, β_1) .

- Posteriorfördelningen för förväntad bensinförbrukning för alla möjliga bilar som väger mellan 0.8 och 2.0 ton.
- Kredibilitetsintervall för \hat{y}_i = kredibilitetsintervall för μ_i som funktion av olika bilvikter x_i i en figur.
- Prediktionsintervall för y_i som funktion av olika bilvikter x_i i en figur.

- Beräkna den klassiska förklaringsgraden för varje samplat värde j från posteriorn:

$$R_j^2 = \frac{SSR_j}{SST} = \frac{\sum_{i=1}^n (\mu_{ij} - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- Gelman et al (2017) argumenterar för att en alternativ förklaringsgrad är bättre, eftersom förklaringsgraden ovan kan leda till $R^2 > 1$ vid lite data och informativa priorfördelningar, se dokument

Bayesian_Rsquare.pdf på Lisam.

- Alternativ Bayesiansk förklaringsgrad:

$$\text{Bayesian } R_j^2 = \frac{SSR_j}{SSR_j + SSE_j} = \frac{\sum_{i=1}^n (\mu_{ij} - \bar{y})^2}{\sum_{i=1}^n (\mu_{ij} - \bar{y})^2 + \sum_{i=1}^n (y_i - \mu_{ij})^2}$$

- Ej standard med residualanalys i Bayesiansk statistik. Val av modell är subjektivt och det är mer vanligt att utvärdera konkurrerande modeller mot varandra utifrån två huvuddrag:
 - 1 hur bra är modellen på att replikera data (*in-sample fit*)
 - 2 hur bra prediktionsförmåga har modellen (*out-of sample fit*).
- Om residualanalys används, så kan den utvärderas på “vanligt” sätt. Obs! För varje observation i data har man en posteriorfördelning över residualen för denna observation: $r_{ij} = y_i - \mu_{ij}$ för varje samplat värde j .
- Plotta residualerna mot μ för att undersöka om det är konstant variation σ kring μ . Undersök även här hur bra det linjära antagandet verkar vara för hur μ är länkat till förklaringsvariabler.
- Undersök om residualerna är normalfördelade med histogram.
- Undersök om residualerna verkar vara beroende av varandra över observationsordning.

- I stort sett alla posteriorutvärderingar kan göras för den multipla linjära regressionsmodellen som för den enkla linjära regressionsmodellen, dvs
 - 1 Tabeller över posteriorresultat för respektive parameters marginella posteriorfördelning, t.ex. medelvärde, standardavvikelse och kredibilitetsintervall.
 - 2 Visualisering av marginella posteriorfördelningen för varje lutningsparameter. Inte viktigt vid kvadratisk approximation eller i fall där posteriorn är lik en multivariat normalfördelning. Fokusera på vissa marginella posteriorfördelningar som visar något avvikande.
 - 3 Oddset för positiv eller negativ lutning kan beräknas för respektive lutningsparametern β_j till en förklaringsvariabel j .
 - 4 Visualisering av bivariata posteriorfördelningar kan vara intressant för att undersöka hur effekten från två förklaringsvariabler samvarierar.
 - 5 Parametrarnas korrelationsmatris kan redovisas.

- I stort sett alla posteriorutvärderingar kan göras för den multipla linjära regressionsmodellen som för den enkla linjära regressionsmodellen, t.ex.
 - 1 Posteriorfördelningen för μ kan redovisas för specifika värden på vektorn med förklaringsvariabler x .
 - 2 Prediktionsintervall och den prediktiva fördelningen för y kan redovisas för specifika värden på vektorn med förklaringsvariabler x .
 - 3 Replikering av data kan jämföras med faktiska data för modellutvärdering. (*in-sample fit*)
 - 4 Prediktioner för nya värden kan redovisas från den prediktiva fördelningen $p(\tilde{y}|y)$. Den prediktiva förmågan kan utvärderas mellan modeller utifrån olika prediktionsmått. (*out-of sample fit*)

- Multipl linjär regressionsanalys med
 - beroende variabel $y = \text{liter} / \text{mil}$ för en bils bränsleförbrukning
 - $x_1 =$ manuell växellåda (=1)
 - $x_2 =$ vikt i ton
 - $x_3 =$ antal hästkrafter
 - $x_4 =$ tid i sek på en kvarts mile
 - $x_5 =$ antal framåtväxlar
- Standardisera alla förklaringsvariabler förutom dummyvariabeln x_1 .

- Modell:

$$Y_1, \dots, Y_n | \mu_i, \sigma^2 \stackrel{iid}{\sim} N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} = \boldsymbol{\beta} \mathbf{x}_i',$$

där variansen σ^2 är okänd och med vektorn av förklaringsvariabler som $\mathbf{x}_i = (1 \ x_{1i} \ \dots \ x_{5i})$ för observation i samt vektorn med parametrar $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_5)$.

- Om $\boldsymbol{\beta}$ och σ antas oberoende apriori (boken):

$$p(\boldsymbol{\beta}, \sigma) = \prod_{j=0}^k p(\beta_j) p(\sigma).$$

Multipel linjär regression - uniform prior $p(\beta, \ln \sigma)$

■ Modell:

$$Y_1, \dots, Y_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$$
$$\mu = \beta \mathbf{x}'$$

■ Prior:

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-1}$$
$$p(\beta | \sigma^2) \propto c$$
$$p(\sigma^2) \propto (\sigma^2)^{-1}$$

■ Betingad posterior:

$$\beta | \sigma^2, y, \mathbf{X} \sim N(\hat{\beta}, V_{\beta} \sigma^2),$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'y$$

$$V_{\beta} = (\mathbf{X}'\mathbf{X})^{-1},$$

där \mathbf{X} är en $n \times (k+1)$ matris med enhetsvektorn och k stycken förklaringsvariabler.

- Marginell posterior för $\sigma^2 | \mathbf{x}, y$:

$$\sigma^2 | y \sim \text{Inv} - \chi^2(n - k - 1, s^2),$$

$$\text{där } s^2 = \frac{(y - \mathbf{x}\beta')'(y - \mathbf{x}\beta')}{n - k - 1}$$

- $\text{Inv} - \chi^2(n - k - 1, s^2)$ är en skalad invers χ^2 -fördelning:

$$\frac{(n - k - 1) s^2}{\sigma^2} | y \sim \chi^2(n - k - 1).$$

- Dra värden från posteriorfördelningen för σ^2 :

- 1 Dra ett värde $f(\sigma^2) = \frac{(n-k-1)s^2}{\sigma^2}$ från $\chi^2(n-k-1)$. (använd funktion `rchisq` i R)
- 2 Beräkna det dragna värdet för σ^2 , givet värdet $f(\sigma^2)$ i punkt 1, enligt:

$$\sigma^2 = \frac{(n-k-1)s^2}{f(\sigma^2)}$$

- 3 Upprepa denna procedur många gånger för att få många dragna värden från posteriorfördelningen för σ^2 .

- Dra värden från den betingade posteriorfördelningen $\beta|\sigma^2, y, \mathbf{X}$:

$$\beta_j|\sigma_j^2, y, \mathbf{X} \sim N(\hat{\beta}, V_{\beta}\sigma_j^2),$$

där j är den j :te samplade dragningen från respektive posteriorfördelning för $\sigma^2|\mathbf{X}, y$ och $\beta|\sigma^2, y, \mathbf{X}$.