CE-092: Extensões de Modelos de Regressão – Banco de questões

Modelos de regressão para dados no intervalo unitário

Observações. As questões abertas marcadas com [**R**] devem ser resolvidas com o computador; as demais podem ser resolvidas sem o computador (calculadora é suficiente). Os conjuntos de dados utilizados estão disponíveis nos pacotes `betareg` e `unitregTMB`. Nas saídas reproduzidas, os valores podem estar arredondados.

1 Questões de múltipla escolha

1.1 Questão

Em qual das situações abaixo um modelo de regressão **beta** (ou outra distribuição contínua no intervalo unitário) é a escolha mais natural para a variável resposta?

a) Número de sementes germinadas entre 25 plantadas em cada canteiro.
b) Proporção da renda familiar gasta com alimentação.
c) Número de peças defeituosas em lotes de 500 peças produzidas.
d) Indicador de sucesso (1) ou fracasso (0) de um tratamento para cada paciente.
e) Tempo (em horas) até a falha de um equipamento.

1.2 Questão

Considere ajustar o modelo linear normal $Y_i = \mathbf{x}_i^\top \beta + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, a uma resposta $Y_i \in (0, 1)$. Assinale a alternativa **INCORRETA**.

a) Os valores ajustados $\hat{Y}_i = \mathbf{x}_i^\top \hat{\beta}$ podem ser menores que 0 ou maiores que 1.
b) A suposição $\mathrm{Var}(Y_i \mid \mathbf{x}_i) = \sigma^2$ é pouco plausível, pois respostas com média próxima de 0 ou de 1 têm menos espaço para variar.
c) A distribuição condicional da resposta pode ser fortemente assimétrica, o que não é compatível com a distribuição normal.
d) Se todas as respostas observadas estão em $(0, 1)$, os valores ajustados por mínimos quadrados necessariamente também estarão em $(0, 1)$.
e) O modelo é incoerente com o suporte da variável observada, e não apenas um problema "estético".

1.3 Questão

Seja $Y \sim \mathrm{Beta}(\mu, \phi)$ na parametrização de Ferrari e Cribari-Neto (2004), com $\mu = 0.5$ e $\phi = 4$. A variância de Y é:

a) 0.0625
b) 0.0500
c) 0.1250
d) 0.0100
e) 0.2500

1.4 Questão

Ainda na parametrização $\mathrm{Beta}(\mu, \phi)$, assinale a alternativa correta.

a) Quando $\phi \to \infty$, $\mathrm{Var}(Y) \to \mu(1 - \mu)$.
b) Quando $\phi \to 0$, $\mathrm{Var}(Y) \to \mu(1 - \mu)$, a relação média–variância da distribuição Bernoulli.
c) A variância de Y não depende de μ, apenas de ϕ.
d) Para μ fixo, quanto maior ϕ, maior a variância de Y.
e) Para ϕ fixo, a variância é máxima quando μ está próximo de 0 ou de 1.

1.5 Questão

Sobre as funções de ligação $g : (0,1) \to \mathbb{R}$ usadas para a média μ_i (logit, probit, complemento log-log – cloglog, e cauchit), assinale a alternativa correta.

a) Apenas a ligação logit garante que $\hat{\mu}_i \in (0,1)$; as demais podem gerar médias fora do intervalo.
b) A ligação cloglog não é simétrica em torno de $\mu = 0.5$: por exemplo, $g^{-1}(0) = 1 - e^{-1} \approx 0.63$, enquanto logit, probit e cauchit têm $g^{-1}(0) = 0.5$.
c) A ligação cauchit tem caudas mais leves que a probit, aproximando-se de 0 e 1 mais rapidamente.
d) A escolha da ligação não altera a verossimilhança maximizada do modelo, apenas a interpretação dos coeficientes.
e) A ligação probit é definida por $g(\mu) = \log\{-\log(1-\mu)\}$.

1.6 Questão

No conjunto `FoodExpenditure`, ajustou-se o modelo beta $\text{logit}(\mu_i) = \beta_0 + \beta_1\, \texttt{income}_i + \beta_2\, \texttt{persons}_i$, obtendo-se $\hat{\beta}_2 = 0.118$ e $\exp(0.118) \approx 1.125$. A interpretação correta de $\hat{\beta}_2$ é:

a) Cada pessoa adicional no domicílio aumenta em 0.118 a proporção média da renda gasta com alimentação.
b) Cada pessoa adicional no domicílio aumenta em 12.5% a proporção média μ_i.
c) Mantida a renda constante, cada pessoa adicional aumenta em aproximadamente 12.5% a razão $\mu_i/(1-\mu_i)$.
d) Mantida a renda constante, cada pessoa adicional aumenta em aproximadamente 12.5% a precisão ϕ.
e) Mantida a renda constante, cada pessoa adicional diminui em 11.8% a razão $\mu_i/(1-\mu_i)$.

1.7 Questão

No pacote `betareg`, o comando `betareg(y ~ x1 + x2 | z1, data = dados)` (ligações padrão) especifica o modelo:

a) $\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 z_{i1}$, com ϕ constante.
b) $\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ e $\log(\phi_i) = \gamma_0 + \gamma_1 z_{i1}$.
c) $\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2} z_{i1}$ (interação), com ϕ constante.
d) $\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ ajustado separadamente para cada valor de z_1.
e) $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ e $\text{logit}(\phi_i) = \gamma_0 + \gamma_1 z_{i1}$.

1.8 Questão

No conjunto `GasolineYield`, `batch` é um fator com 10 níveis e `temp` é numérica. Quantos parâmetros são estimados no modelo `betareg(yield ~ batch + temp | temp, data = GasolineYield)`?

a) 11
b) 12
c) 13
d) 14
e) 22

1.9 Questão

No modelo `betareg(yield ~ batch + temp | temp, data = GasolineYield)` obteve-se, para a estrutura de precisão (ligação log), $\hat{\gamma}_1 = 0.0146$ associado a `temp`. Assinale a interpretação correta.

a) Um aumento de 10 unidades na temperatura aumenta a média μ_i em 0.146.
b) Um aumento de 10 unidades na temperatura aumenta ϕ_i em 0.146 unidades.
c) Um aumento de 10 unidades na temperatura multiplica ϕ_i por $e^{0.146} \approx 1.16$; para uma mesma média, isso corresponde a menor variabilidade condicional.
d) Um aumento de 10 unidades na temperatura multiplica a variância de Y_i por $e^{0.146} \approx 1.16$.
e) Como a ligação é log, o coeficiente não admite interpretação.

1.10 Questão

Para o `GasolineYield`, os modelos com precisão constante (`fit_gy1`, 12 parâmetros) e com precisão dependendo de `temp` (`fit_gy2`, 13 parâmetros) apresentaram log-verossimilhanças 84.80 e 86.98, respectivamente. A estatística do teste da razão de verossimilhanças e sua distribuição de referência são:

a) 2.18, com distribuição χ^2_1.

b) 4.36, com distribuição χ^2_1.
c) 4.36, com distribuição χ^2_{13}.
d) 2.18, com distribuição χ^2_{12}.
e) O teste não pode ser feito, pois os modelos têm estruturas de precisão diferentes.

1.11 Questão

Para o conjunto `wqi`, ajustaram-se cinco modelos com a **mesma** estrutura sistemática (`y ~ location + quarter`, ligação logit, precisão constante), mudando apenas a distribuição condicional:

Modelo	logLik	AIC	BIC
Beta	224.62	-435.25	-412.52
Unit-Gamma	227.79	-441.57	-418.84
Simplex	232.44	-450.88	-428.15
Bessel	228.39	-442.77	-420.04
Vasicek	227.65	-441.29	-418.56

Assinale a alternativa correta.

a) Deve-se escolher o modelo beta, por ser a referência clássica para dados em $(0, 1)$.
b) O modelo simplex é o preferido; como os cinco modelos têm o mesmo número de parâmetros, log-verossimilhança, AIC e BIC levam à mesma ordenação.
c) Os modelos não podem ser comparados por AIC, pois não são aninhados.
d) O modelo Vasicek é o preferido, pois tem o menor valor de logLik.
e) O AIC favorece o simplex, mas o BIC favorece o beta.

1.12 Questão

Sobre as distribuições alternativas à beta disponíveis no `unitregTMB` (unit-gamma, simplex, Bessel e Vasicek), todas parametrizadas pela média $\mu = E(Y)$, assinale a alternativa correta.

a) Para os mesmos valores de μ e ϕ, as cinco distribuições têm a mesma variância.
b) Se duas distribuições têm a mesma média e a mesma variância, então têm a mesma densidade.
c) Nas distribuições simplex e Vasicek, ϕ é um parâmetro de **dispersão**: aumentar ϕ aumenta a variabilidade, ao contrário da beta.
d) Ao trocar a distribuição, o `unitregTMB` também troca a função de ligação, o que torna os coeficientes incomparáveis.
e) Trocar a distribuição altera o significado de μ_i, que deixa de ser a média condicional.

1.13 Questão

No modelo misto do `unitregTMB` com distribuição Vasicek, utiliza-se ligação **logit** para o parâmetro ϕ_{ij}, enquanto para as distribuições beta, unit-gamma, simplex e Bessel utiliza-se ligação log. O motivo é:

a) A ligação logit produz estimativas mais precisas que a ligação log.
b) Na Vasicek, $\phi_{ij} \in (0, 1)$, enquanto nas demais $\phi_{ij} > 0$.
c) Na Vasicek, ϕ_{ij} representa a média da distribuição.
d) A ligação log não é diferenciável, impedindo o uso de diferenciação automática.
e) Na Vasicek, ϕ_{ij} pode assumir valores negativos.

1.14 Questão

No conjunto `periodontal`, a resposta é a proporção de sítios dentários com doença, com 9.8% de zeros exatos (ausência de doença) e 8.1% de uns exatos (todos os sítios comprometidos). A abordagem mais adequada, entre as discutidas na disciplina, é:

a) Ajustar diretamente uma regressão beta, pois o suporte da distribuição beta é $[0, 1]$.
b) Eliminar as observações iguais a 0 e a 1 e ajustar um modelo beta às demais.
c) Usar um modelo aumentado (inflacionado) em zero e um: massas pontuais $p_0 = P(Y = 0)$ e $p_1 = P(Y = 1)$, e uma distribuição contínua em $(0, 1)$ com probabilidade $1 - p_0 - p_1$.
d) Usar regressão logística para a variável indicadora de $Y > 0.5$.
e) Usar o modelo linear normal, que não tem restrição de suporte.

1.15 Questão

Em um modelo aumentado em zero e um, para certo perfil de covariáveis tem-se $p_0 = P(Y = 0) = 0.10$, $p_1 = P(Y = 1) = 0.08$ e a componente contínua (em $(0, 1)$) tem média $\mu_c = 0.40$. O valor de $E(Y)$ é:

a) 0.400
b) 0.408
c) 0.480
d) 0.328
e) 0.388

1.16 Questão

O conjunto `wqi` contém o IQA (reescalado em $(0, 1)$) medido em 16 usinas, em 3 locais (montante, reservatório e jusante) e 4 trimestres, totalizando 190 observações (há dois dados faltantes). Na notação y_{ij} do modelo misto, com i a unidade amostral e j a observação dentro da unidade, tem-se:

a) $i = 1, \dots, 190$ e $j = 1, \dots, 16$.
b) $i = 1, \dots, 16$ (usinas) e $j = 1, \dots, n_i$, com $n_i = 12$ para a maioria das usinas e $\sum_i n_i = 190$.
c) $i = 1, \dots, 3$ (locais) e $j = 1, \dots, 4$ (trimestres).
d) $i = 1, \dots, 4$ (trimestres) e $j = 1, \dots, 48$.
e) $i = 1, \dots, 16$ e $j = 1, \dots, 12$, com $n_i = 12$ para todas as usinas.

1.17 Questão

Considere o modelo misto $\text{logit}(\mu_{ij}) = \mathbf{x}_{ij}^\top \beta + u_i$, $u_i \sim N(0, \sigma^2)$. Assinale a alternativa correta.

a) σ^2 mede a variabilidade entre as observações de uma mesma unidade.
b) Os u_i são parâmetros fixos, estimados um para cada unidade, como se `id` fosse incluído como fator na fórmula.
c) σ^2 mede a heterogeneidade entre as unidades; se $\sigma^2 = 0$, o modelo se reduz ao modelo com observações independentes.
d) O intercepto aleatório entra no preditor linear da precisão ϕ_{ij}.
e) O intercepto aleatório induz dependência apenas entre observações de unidades diferentes.

1.18 Questão

Por que a aproximação de Laplace é necessária na estimação dos modelos mistos para dados no intervalo unitário?

a) Porque a verossimilhança condicional $\prod_j f_c(y_{ij}; \mu_{ij}, \phi_{ij})$ não pode ser avaliada numericamente.
b) Porque a verossimilhança marginal exige integrar os efeitos aleatórios, integral que em geral não tem solução analítica; a aproximação usa uma expansão quadrática da log-verossimilhança completa em torno da moda condicional $\hat{\mathbf{u}}_i$.
c) Porque a aproximação de Laplace é exata para qualquer distribuição da resposta.
d) Porque é preciso integrar a verossimilhança em relação aos efeitos fixos β.
e) Porque a aproximação de Laplace evita a necessidade de estimar σ.

1.19 Questão

Sobre a implementação dos modelos mistos no pacote `unitregTMB` via *Template Model Builder* (TMB), assinale a alternativa **INCORRETA**.

a) A log-verossimilhança completa é escrita em C++.
b) Para cada valor de θ, o TMB determina internamente a moda condicional $\hat{\mathbf{u}}_i$.
c) As derivadas da função objetivo são obtidas por diferenciação automática.
d) Erros-padrão (inclusive de transformações não lineares dos parâmetros) são obtidos com `TMB::sdreport()`.
e) Os efeitos aleatórios são estimados primeiro por mínimos quadrados e depois tratados como covariáveis fixas na maximização da verossimilhança.

1.20 Questão

Ajustou-se o modelo `unitregTMB(y ~ x + (1 | id), family = kumaraswamy(model_for = "quantile", tau = 0.1), data = dados)`. Assinale a alternativa correta.

a) μ_{ij} é a média das 10% menores observações da unidade i.

b) μ_{ij} é o quantil de ordem 0.1 da distribuição condicional de Y_{ij} (dados $\mathbf{x}_{ij}$ e u_i), e $\tau = 0.1$ é fixado pelo usuário, não estimado.
c) τ é estimado junto com β e ϕ por máxima verossimilhança.
d) Com $\tau = 0.1$, o modelo descreve 10% da variância de Y.
e) Com `model_for = "quantile"`, μ_{ij} continua sendo a média condicional $E(Y_{ij} \mid u_i)$.

1.21 Questão

No exemplo do percentual de gordura corporal (`bodyfat_long`), o modelo misto beta com precisão constante (Beta1) tem log-verossimilhança 2198.67 e 12 parâmetros; o modelo beta com precisão variável (Beta2), em que $\log(\phi_{ij}) = \gamma_0 + \gamma_1 \text{IMC}_i + \gamma_{2j}\text{região}_j + u_{i2}$, tem log-verossimilhança 2303.41. O teste da razão de verossimilhanças entre Beta1 e Beta2 tem estatística e graus de liberdade:

a) 104.74 e 6.
b) 209.48 e 6.
c) 209.48 e 5.
d) 209.48 e 18.
e) 104.74 e 1.

1.22 Questão

No modelo unit-gama misto com precisão variável para o percentual de gordura corporal, a estrutura de precisão (ligação log, região de referência: braços) apresentou $\hat{\gamma}_{25} = 2.80$ para a região ginoide e $\hat{\gamma}_{24} = -0.94$ para a região androide. Assinale a alternativa correta.

a) A gordura da região ginoide tem média $e^{2.80} \approx 16.4$ vezes a dos braços.
b) Mantidas as demais covariáveis, a precisão na região ginoide é cerca de $e^{2.80} \approx 16.4$ vezes a dos braços, isto é, menor variabilidade condicional; na região androide, a precisão é menor que a dos braços.
c) A região androide é a de menor percentual médio de gordura, pois $\hat{\gamma}_{24} < 0$.
d) A variância na região ginoide é 2.80 unidades menor que a dos braços.
e) Os coeficientes γ não são interpretáveis, pois o modelo tem efeito aleatório.

1.23 Questão

Para o `wqi` com intercepto aleatório por usina, comparou-se o modelo beta com precisão constante (`fit_beta`) e com precisão dependendo de `quarter + location` (`fit2_beta`):

```
anova(fit_beta, fit2_beta)
          Df     AIC     BIC logLik   Chisq Chi Df Pr(>Chisq)
fit_beta   8 -446.09 -420.12 231.05
fit2_beta 13 -446.54 -404.33 236.27 10.4450      5    0.06356
```

Assinale a conclusão mais adequada.

a) A precisão variável é claramente necessária, pois a log-verossimilhança aumentou.
b) Ao nível de 5%, não se rejeita a hipótese de precisão constante; o BIC também favorece `fit_beta` e o AIC é praticamente igual nos dois — prefere-se o modelo mais parcimonioso.
c) Os modelos não são aninhados, logo o teste da razão de verossimilhanças não é válido.
d) O teste tem 13 graus de liberdade.
e) O AIC e o BIC sempre levam à mesma conclusão, por isso basta olhar um deles.

1.24 Questão

No modelo misto com ligação logit $\text{logit}(\mu_{ij}) = \mathbf{x}_{ij}^\top \beta + u_i$, $u_i \sim N(0, \sigma^2)$, $\sigma > 0$, assinale a alternativa correta sobre a interpretação dos efeitos fixos.

a) $\text{logit}^{-1}(\mathbf{x}_{ij}^\top \beta)$ é exatamente a média marginal (populacional) $E(Y_{ij})$.
b) $\text{logit}^{-1}(\mathbf{x}_{ij}^\top \beta)$ é a média condicional para uma unidade "típica", com $u_i = 0$; por causa da não linearidade da ligação, ela em geral não coincide com a média marginal $E(Y_{ij})$.
c) Os coeficientes β são iguais aos do modelo sem efeito aleatório, por construção.
d) e^{β_k} é a razão entre as médias μ_{ij} para um aumento unitário em x_k.
e) Os efeitos fixos só podem ser interpretados se $\sigma^2 = 0$.

2 Questões abertas

2.1 Questão

Considere o conjunto `FoodExpenditure` (pacote `betareg`), com a resposta `prop = food/income`, e a saída a seguir.

```
fit_beta <- betareg(prop ~ income + persons, data = FoodExpenditure)
summary(fit_beta)
Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.622548   0.223854  -2.781 0.005418
income      -0.012299   0.003036  -4.052 5.09e-05
persons      0.118462   0.035341   3.352 0.000802
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    35.61       8.08   4.407 1.05e-05
Log-likelihood: 45.33 on 4 Df
```

a) Escreva o modelo ajustado (distribuição, preditor linear, função de ligação e estrutura de precisão).
b) Calcule a proporção média estimada da renda gasta com alimentação para uma família com `income = 50` e `persons = 4`.
c) Calcule a variância estimada de Y para essa mesma família.
d) Interprete os efeitos de `income` e `persons` em termos de $\exp(\hat{\beta})$.
e) Para esses dados, o modelo linear `lm(prop ~ income + persons)` produziu valores ajustados entre 0.153 e 0.455, todos dentro de $(0, 1)$. Isso torna o modelo linear adequado? Justifique.

2.2 Questão [R]

Utilize o conjunto `GasolineYield` do pacote `betareg` (`data("GasolineYield", package = "betareg")`), com a resposta `yield` (proporção de petróleo bruto convertido em gasolina), o fator `batch` e a temperatura `temp`.

a) Ajuste o modelo beta com precisão constante, `yield ~ batch + temp`, e o modelo com precisão variável, `yield ~ batch + temp | temp`. Escreva as equações de ambos e informe o número de parâmetros de cada um.
b) Compare os dois modelos pelo teste da razão de verossimilhanças, pelo AIC e pelo BIC. Qual modelo você adotaria?
c) Interprete a estimativa do coeficiente de `temp` na estrutura de precisão.
d) Para o lote `batch = 1`, obtenha a média, a precisão e a variância estimadas de `yield` para `temp` = 250, 300 e 400. Comente o comportamento da média e da variância.
e) Reajuste o modelo de precisão variável com as ligações `probit`, `cloglog` e `loglog` para a média (argumento `link`). Qual ligação produz o melhor ajuste?

2.3 Questão [R]

Ainda com o `FoodExpenditure` (resposta `prop = food/income`):

a) Ajuste o modelo beta `prop ~ income + persons` (precisão constante) e o modelo em que a precisão depende do número de pessoas, `prop ~ income + persons | persons`.
b) Teste, pela razão de verossimilhanças, se a precisão é constante. Apresente a estatística, os graus de liberdade e o valor-p.
c) Interprete o coeficiente de `persons` na estrutura de precisão.
d) As estimativas dos coeficientes de média mudaram de um modelo para o outro? Comente.

2.4 Questão [R]

O conjunto `ReadingSkills` (pacote `betareg`) contém escores de precisão de leitura (`accuracy`) de 44 crianças, com e sem dislexia (`dyslexia`), e o QI padronizado (`iq`).

a) Ajuste o modelo linear `lm(accuracy ~ dyslexia * iq)` e o modelo beta `betareg(accuracy ~ dyslexia * iq | dyslexia + iq)`. Escreva o modelo beta.
b) Compare o modelo beta acima com o modelo beta de precisão constante (`accuracy ~ dyslexia * iq`). Qual é preferido?
c) Obtenha a média e a precisão estimadas pelo modelo beta para crianças com e sem dislexia com `iq` igual a -1, 0 e 1. Descreva o efeito do QI em cada grupo.

d) Compare as predições do item (c) com as do modelo linear. Em que situação as diferenças são maiores e por quê?

2.5 Questão [R]

Considere o conjunto `wqi` do pacote `unitregTMB` (IQA de 16 usinas, medido em 3 locais e 4 trimestres).

a) Ajuste o modelo beta `y ~ quarter + location` (observações independentes) e o modelo beta misto `y ~ quarter + location + (1 | id)`, ambos com precisão constante (`family = beta_fam`). Escreva o modelo misto, identificando os índices i e j.
b) Compare os dois modelos com `anova()` e com `gof_tab()`. A inclusão do intercepto aleatório é justificada? (Comente sobre a hipótese testada.)
c) Informe e interprete a estimativa do desvio-padrão do intercepto aleatório.
d) Use `random_effects()` para identificar a usina com maior e com menor efeito aleatório predito.
e) Calcule a média condicional estimada do IQA no reservatório, no 3º trimestre, para (i) uma usina típica ($u_i = 0$) e (ii) a usina com maior efeito predito.
f) Ajuste o mesmo modelo misto com as famílias `simplex`, `unitgamma`, `bessel` e `vasicek` e indique qual distribuição se ajusta melhor.

2.6 Questão

Para o conjunto `wqi`, foi ajustado o modelo misto `unitregTMB(y ~ quarter + location + (1 | id), family = beta_fam, data = wqi)` com a saída (resumida):

```
Mu coefficients (logit link):
                   Estimate Std. Error
(Intercept)         1.13560    0.08833
quarter2            0.21886    0.09010
quarter3            0.30926    0.09077
quarter4            0.05210    0.08766
locationReservoir   0.23701    0.07817
locationDownstream  0.16363    0.07776
Phi coefficients (log link):
(Intercept)         3.4169     0.1067
Random Effects in Mu:  Groups: 'id' (16 levels)
    (Intercept)  Std. Deviation 0.1858   (Std. Error 0.049)
Log-Likelihood: 231.05
```

a) Escreva o modelo ajustado, indicando os parâmetros estimados e o significado de cada índice.
b) Calcule $\widehat{\phi}$ e a média condicional estimada do IQA no reservatório, no 3º trimestre, para uma usina típica ($u_i = 0$).
c) Calcule a variância condicional estimada da resposta nessa mesma situação.
d) Calcule o intervalo em que se espera encontrar a média condicional de 95% das usinas nessa mesma situação.
e) Explique por que $\text{logit}^{-1}(\mathbf{x}^\top \widehat{\beta})$ calculado em (b) não é a média marginal do IQA nessa condição. A média marginal seria maior ou menor?

2.7 Questão

Considere um modelo misto para dados no intervalo unitário com $Y_{ij} \mid u_i \sim \mathcal{D}(\mu_{ij}, \phi)$, $\text{logit}(\mu_{ij}) = \mathbf{x}_{ij}^\top \beta + u_i$, $u_i \sim N(0, \sigma^2)$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$.

a) Escreva a verossimilhança condicional da unidade i e a verossimilhança marginal do modelo.
b) Explique por que a verossimilhança marginal não pode, em geral, ser calculada analiticamente e descreva a aproximação de Laplace (moda condicional, hessiana e expressão aproximada de $\ell_i^*(\theta)$).
c) Descreva as etapas da estimação no `unitregTMB` (TMB).
d) Uma alternativa ao efeito aleatório seria incluir `id` como fator (efeito fixo). Discuta vantagens e desvantagens de cada abordagem para o `wqi` (16 usinas) e para o `bodyfat_long` (298 indivíduos).

2.8 Questão

No estudo de gordura corporal (`bodyfat_long`, 298 indivíduos $\times$ 5 regiões), foram ajustados os modelos a seguir (UG = unit-gama; 1: precisão constante; 2: precisão variável com $\log(\phi_{ij}) = \gamma_0 + \gamma_1 \text{IMC}_i + \gamma_{2j} \text{região}_j + u_{i2}$). Todos têm intercepto aleatório $u_{i1} \sim N(0, \tau_1^2)$ na média.

Parâmetro	UG1	Beta1	UG2	Beta2
b0: Intercepto	-2.806(0.122)	-2.815(0.122)	-2.179(0.127)	-2.187(0.127)
b1: Idade	0.002(0.001)	0.002(0.001)	-0.003(0.010)	-0.003(0.010)
b2: IMC	0.084(0.005)	0.084(0.005)	0.066(0.006)	0.066(0.006)
b3: Sexo-masculino	-0.644(0.030)	-0.650(0.030)	-0.757(0.030)	-0.760(0.030)
(b4: IPAQ e b5j: Região -- 6 coeficientes omitidos)				
phi	56.716(2.330)	96.136(3.945)	-	-
tau1	0.226(0.011)	0.226(0.011)	0.244(0.011)	0.244(0.011)
g0: Intercepto	-	-	2.730(0.413)	3.152(0.418)
g1: IMC	-	-	0.048(0.016)	0.052(0.016)
(g2j: Região -- 4 coeficientes omitidos)				
tau2	-	-	0.173(0.104)	0.186(0.099)
LogLik	2208.64	2198.67	2305.82	2303.41
AIC	-4393.28	-4373.34	-4575.64	-4570.82

a) Escreva o modelo UG2 completo (distribuição, preditores para μ e ϕ, efeitos aleatórios).
b) Qual dos quatro modelos é preferido? Justifique.
c) Compare UG1 e UG2 pelo teste da razão de verossimilhanças (determine os graus de liberdade).
d) Interprete o efeito do sexo masculino na média, no modelo UG2.
e) Os valores de $\hat{\phi}$ em UG1 e Beta1 são muito diferentes (56.7 e 96.1). Isso indica que os modelos discordam quanto à variabilidade dos dados? Explique.
f) Com base nos resultados, que simplificação do modelo UG2 você sugeriria?

2.9 Questão

No estudo da doença periodontal (`periodontal`, 290 indivíduos $\times$ 4 tipos de dentes), Y_{ij} é a proporção de sítios com doença, com 9.8% de zeros e 8.1% de uns. Ajustou-se um modelo misto unit-Weibull aumentado em zero e um, para a mediana ($\tau = 0.5$), com p_0 e p_1 constantes:

```
unitregTMB(y ~ gender + age + hba1c + smoker + tooth + (1 | id),
           family = unitweibull(model_for = "quantile", tau = 0.5), data = periodontal)
Mu coefficients (logit link):
                  Estimate         Zero-inflation (p0):  (Intercept) -2.12237
(Intercept)       -2.915934        One-inflation  (p1):  (Intercept) -2.31530
genderFemale      -0.613181
age                0.038437        Random Effects in Mu: Std. Deviation 1.216
hba1cUncontrolled  0.098843        Log-Likelihood: -372.58  (12 parâmetros)
smokerSmoker       0.143904
toothMolar         2.362204     (categorias de referência: Male, Controlled,
toothPremolar      0.922617      Non-smoker, Canine)
toothIncisor       0.151983
```

Para p_0 e p_1 o pacote usa a ligação logit "bivariada" (multinomial): $p_0 = \dfrac{e^{\eta_0}}{1 + e^{\eta_0} + e^{\eta_1}}$ e $p_1 = \dfrac{e^{\eta_1}}{1 + e^{\eta_0} + e^{\eta_1}}$.

a) Descreva a estrutura do modelo para Y_{ij} (as três partes: $Y = 0$, $0 < Y < 1$, $Y = 1$) e escreva a densidade/probabilidade de $Y_{ij} \mid u_i$.
b) Calcule $\hat{p}_0$, $\hat{p}_1$ e $P(0 < Y < 1)$. Compare com as proporções observadas.
c) Calcule a mediana condicional estimada da parte contínua para um homem de 50 anos, HbA1c controlado, não fumante, com $u_i = 0$, para (i) caninos e (ii) molares.
d) Interprete $\exp(2.362)$ e $\exp(-0.613)$.
e) Um segundo modelo, com todas as covariáveis em p_0 e p_1, teve log-verossimilhança -230.79 com 26 parâmetros. Compare os dois modelos.
f) Por que modelar quantis (e não a média) pode ser interessante neste problema?

2.10 Questão

Discuta, de forma comparativa, as seguintes extensões do modelo de regressão beta com precisão constante, $Y_i \sim \text{Beta}(\mu_i, \phi)$, $\text{logit}(\mu_i) = \mathbf{x}_i^\top \beta$:

(1) precisão variável; (2) outras distribuições unitárias (unit-gama, simplex, Bessel, Vasicek); (3) efeitos aleatórios; (4) modelos aumentados em zero e/ou um; (5) modelos para quantis.

Para cada uma: a) indique qual limitação do modelo de partida ela resolve; b) indique como ela é especificada em R (`betareg`/`unitregTMB`); c) indique como você decidiria, com base nos dados, se ela é necessária.