

MA 575 – Qualifying Exam

Spring 2011

- We are given n data observations of response Y and predictors X and, after fitting a linear model $E[Y|X] = X\boldsymbol{\beta}$ with p parameters *including the intercept*, we decide to apply a Box-Cox transformation to improve the fit.

Recall that a Box-Cox transformation with power λ of Y is given by

$$\Psi(Y, \lambda) = \begin{cases} (Y^\lambda - 1)/\lambda \cdot G^{1-\lambda} & \text{if } \lambda \neq 0, \\ \log Y \cdot G & \text{if } \lambda = 0, \end{cases}$$

where G is the geometric mean of Y . The new model is then given by

$$E[\Psi(Y, \lambda) | X] = X\boldsymbol{\beta}. \quad (*)$$

Since n is large, we will be following a method proposed by Atkinson to fit λ based on a *first-order* approximation of $\Psi(Y, \lambda)$ around $\lambda = 1$.

- (a) Since

$$\Psi(Y, \lambda) \approx \Psi(Y, 1) + \Psi'(Y, 1)(\lambda - 1) = Y - 1 + (\lambda - 1)U(Y),$$

where $U(Y) = Y \log Y + (1 - Y)(1 + \log G)$, show that model (*) can be approximated by

$$E[Y|X, U] = X\boldsymbol{\beta} + U\gamma, \quad (**)$$

if we assume that U is just another predictor. What is the interpretation of γ ?

- Atkinson's approximation is well suited for an *added-variable* model of the residuals of Y given X against the residuals of U given X . Let H be the hat matrix from the first, original linear fit, \mathbf{y} the vector of responses, and \mathbf{u} the vector of U predictors. Provide a least-squares estimate $\hat{\lambda}$ for λ , based on model (**), as a function of H , \mathbf{y} , and \mathbf{u} .
- If $\hat{\sigma}_0$ is the residual standard error for the added-variable model, find the residual standard error for the full model (**). (*Hint*: note that the residuals from the added-variable model are the same as the residuals from the full model.)
- Provide a concise expression for the standard error of $\hat{\lambda}$ as a function of $\hat{\sigma}_0$, \mathbf{u} , and H . Now construct a *t-test* for **no** transformation of the response under the full model: specify the hypothesis, the test statistic, and its distribution under the null.
- Propose a *graphical* procedure to spot observations that could artificially suggest or mask the need for transforming the response.

- Three groups of n observations are fitted using the following fixed-effects model:

$$y_{ij} = \mu + \theta_i + e_{ij}, \quad e_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2),$$

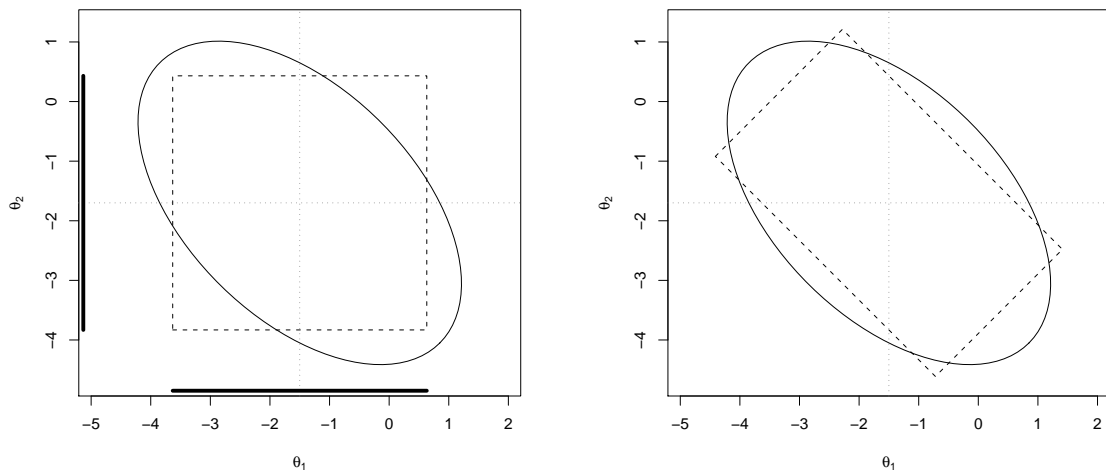
for $i = 1, 2, 3$ and $j = 1, \dots, n$. To avoid identifiability issues, we set $\sum_{i=1}^3 \theta_i = 0$ and remove θ_3 from the above formulation, that is, the parameters in our model are μ , θ_1 , and θ_2 .

- (a) Specify the design matrix and compute the *correlation* between the least squares estimates (LSE) $\hat{\theta}_1$ and $\hat{\theta}_2$.
- (b) After fitting the model, you obtain LSE $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\sigma}^2$. Construct $100(1 - \alpha)\%$ confidence intervals for θ_1 and θ_2 based on these estimates.
- (c) Show that a $100(1 - \alpha)\%$ *joint* confidence region for θ_1 and θ_2 can be specified by

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \frac{2\hat{\sigma}^2}{n} F(\alpha; 2, 3(n - 1)),$$

where $\boldsymbol{\theta} = [\theta_1 \ \theta_2]^T$, and $F(\alpha; 2, 3(n - 1))$ is the $1 - \alpha$ quantile of an $F_{2,3(n-1)}$ distribution.

- (d) The confidence regions from the two previous items are pictured below in figure (a): the bold lines mark the separate confidence intervals, the dashed square is the (Cartesian) product of these intervals, and the ellipsoid represents the joint confidence region.



(a)

(b)

If you were to test the hypothesis of no difference across groups, what would be your conclusions from figure (a) using: (i) the separate confidence intervals and (ii) the joint confidence region? Explain why your conclusions from (i) and (ii) are not consistent.

- (e) Not happy with the separate confidence intervals above, you decide to reparametrize the model using $\gamma_1 = (\theta_1 + \theta_2)/2$ and $\gamma_2 = (\theta_1 - \theta_2)/2$. The product of $100(1 - \alpha)\%$ confidence intervals for γ_1 and γ_2 are represented by the dashed rectangle in figure (b). If you now decide to test the hypothesis of no difference across groups using these new CI, what is your conclusion? Explain why your new conclusion differs from (i) above.