- estimation of λ: choose λ to fit data well using *maximum likelihood*.
  - can compute $L(\lambda)$ for various values of λ and compute $\hat{\lambda}$ exactly to maximize $L(\lambda)$ But, it's ok for prediction purpose $\hat{\lambda}_{MLE}$
  - but usually $\hat{\lambda}$ is not a nice round number, e.g., $\hat{\lambda} = -0.17$. It would be hard to explain what this new response means.
  - to avoid this, maximize $L(\lambda)$ over a grid of values, such as $\{2, 1, 1/2, 0, -1/2, -1, -2\}$. This helps with interpretation Y_λ $\hat{\beta}_\lambda$
  - for $\hat{\lambda}$ outside $[-2, 2]$, pay more attention on whether such transformation is required

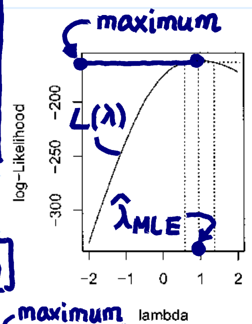
∵ Box-Cox transformation is easy to get upset by outliers

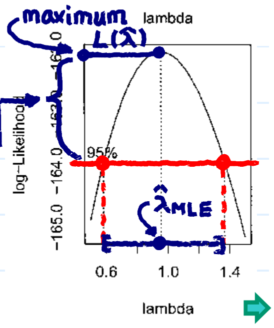
  - **Q**: why not just minimize RSS_λ to estimate λ? Note. RSS_λ has different unit & scale for different λ.
- test of λ: is the transformation really necessary?
  - we can answer the question form a C.I. for λ
  - *likelihood ratio test* (H_0: $\lambda = \lambda_0$ vs. H_A: $\lambda \neq \lambda_0$):

    $-2[L(\lambda_0) - L(\hat{\lambda})] \sim \chi_1^2$ under H_0 $\hat{\lambda}_{MLE}$ $\dim(H_0 \cup H_A) - \dim(H_0)$

    accept H_0 if $-2[\mathcal{L}(\lambda_0) - \mathcal{L}(\hat{\lambda})] \leq \chi_1^2(1-\alpha)$
  - a $100(1-\alpha)\%$ C.I. for λ can be formed by:

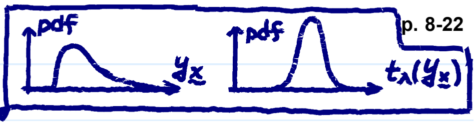
    $\{\lambda \mid L(\lambda) > L(\hat{\lambda}) - (1/2)\,\chi_1^2(1-\alpha)\}$
  - is $\lambda=1$ in the C.I.? if so, may as well stay with no transformation.
  - if rounding $\hat{\lambda}$, check that rounded value is in the C.I.





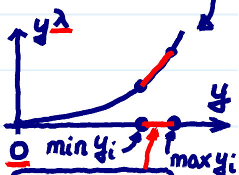
➢ some notes:

- the Box-Cox method gets upset by outliers --- e.g., if see $\hat{\lambda}=5$, this is probably the reason (**Q**: why?)



apply Box-Cox transformation on

- what if some $y_i<0$? adding a constant $\Rightarrow y_i^* = y_i + c$ s.t. $y_i^* > 0$ for all i (c cannot
- if $\max_i y_i/\min_i y_i$ (> 1) is small, Box-Cox won't do anything --- power be too large) transforms are well approximated by linear transformations over short intervals
- should the estimation of λ count as an extra parameter to be taken account of in the degrees of freedom? --- difficult question



Recall. asymmetric normal prob. plot (LNp.7-16)

- Box and Cox (1964) formulate the problem of choosing λ to *make the errors as nearly like a normal sample as possible* $t_\lambda(y_x)$ odd mean $y^\lambda \approx ay+b$

• transformation of predictors ⇐ use different base functions ⇐ Recall. LNp.7-14, remedy for curvature in residual plot

➢ Recall: can use some graphical methods, such as added variable plots and partial residual plot, to offer suggestions for transforming the predictors

➢ could consider Box-Cox family of transformation for each predictor as follows:

$$y=\beta_0+\beta_i x_i+\Sigma_{j\neq i}\,\beta_j x_j+\varepsilon \Rightarrow y=\beta_0+\beta_i^*\,t_\lambda(x_i)+\Sigma_{j\neq i}\,\beta_j x_j+\varepsilon,$$ where t_λ can be x^λ or $log(x)$

- pick λ that minimizes RSS_λ (**Q**: why only RSS_λ here?) to transform x_i to $t_\lambda(x_i)$
- repeat the procedure for each $i \Rightarrow$ lot of works cf. Why cannot just minimize RSS_λ in LNp.8-21?
- correct transformation for each predictor may depend on getting the others right $\Rightarrow$ may need to perform the procedure for all i's several rounds

➢ a simpler method

known functions

can do it for all x_i's simultaneously

- approximate x_i^λ by $x_i + (\lambda-1)\, x_i \log(x_i)$ (i.e., first 2 terms in Taylor's expansion of x^λ w.r.t. λ) to determine the best $\lambda \Rightarrow$ add the terms $x_i\log(x_i)$ to this model

at $\lambda=1$

- suppose $x_i\log(x_i)$ has regression coefficient $\eta \;\Rightarrow\;$ test H_0: $\eta=0$. $\leftarrow \eta = \beta_i^*(\lambda-1)$, $\eta = 0$, $\lambda = 1$
  If accept, no transformation; if rejected, do transformation
- $\beta_i^* x_i^\lambda \approx \beta_i^* [\, x_i + (\lambda-1)\, x_i \log(x_i)\,] \;\Rightarrow\; \hat{\eta} = \hat{\beta}_i^*(\lambda-1) \;\Rightarrow\; \hat{\lambda} = (\hat{\eta}\,/\hat{\beta}_i^*)+1$

from $x_i\log(x_i)$ from x_i

- Some issues in transformation
  ➢ transformation can be used to e.g., better prediction
    - stabilize variance ← e.g., LNp.7-11, table
    - improve fitting ← e.g. LNp.7-15, table
    - make errors nearly normally distributed ← e.g. Box-Cox transformation
    - a transformation of scale may also allow use of a simpler model ← discussion in LNp.7-14
  ➢ these four goals for transformation will not always be met by the same transformation, and compromises may be required
  ➢ transformation of Y can alter the error structure, e.g., additive ↔ multiplicative in exp/log. In practice, try different transformation and check if the residuals satisfy the conditions required for linear regression

cf. $t_\lambda(x_i)$

$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \varepsilon$
$\rightarrow \log(y) = \log(\beta_0) + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \log(\varepsilon)$

  ➢ prediction in Y-space $\Rightarrow$ back-transforming, same for C.I. for the prediction of Y ← C.I. of Y_λ
  ➢ It may be difficult to relate the parameters of the untransformed model to the parameters of transformed model. After transforming, regression coefficients will need to interpreted w.r.t. the transformed scale.

e.g. $y = \beta_0 + \beta_1 x + \varepsilon$
$t_\lambda(y) = \beta_0' + \beta_1' x + \varepsilon$

❖ **Reading**: Faraway (1st ed.), 7.1 ❖ **Further reading**: D&S, chapters 13