STAT 425 — Spring 2016 — Sections 1UG, 1GR

Homework 4

Please submit your assignment *on paper*, following the Formatting Guidelines for Homework Submission. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. Consider the data set `lathe1` from package `alr4`.

   (a) [2 pts] Fit a full *second-order polynomial* model of the response `Life` versus the variables `Feed` and `Speed`. That is, your model should include both quadratic terms and the interaction term. Produce an R summary of your fitted model. Does the interaction term appear to be statistically significant?

   (b) [2 pts] Produce the usual R diagnostic plots for your model from the previous part.

   (c) [2 pts] Briefly assess the diagnostic plots.

   (d) [2 pts] Using the `boxcox` function from the package `MASS`, produce a plot of the (profile) Box-Cox log-likelihood, versus λ.

   (e) [2 pts] Approximately what λ value appears to be selected by the Box-Cox procedure? (Hint: `boxcox(..., plotit=FALSE)` shows you the λ (`x`) values and log-likelihood (`y`) values used in the plot.)

   (f) [2 pts] Find the most "simple" λ value that is still within the confidence limits shown on your plot from part (d). To what kind of a "simple" transformation (i.e. what function) does this correspond?

   (g) [2 pts] Using your "simple" transformation from the previous part, refit the model. Produce an R summary of your fitted model. Does the interaction term appear to be statistically significant?

   (h) [2 pts] Produce the usual R diagnostic plots for your model from the previous part. Have they improved?

2. The data set `trees`[1] contains the `Girth` (diameter, inches), `Height` (feet), and `Volume` (cubic feet) of timber in 31 felled black cherry trees. Natural physical considerations would suggest that

$$\texttt{Volume} \propto (\texttt{Girth})^2 \cdot \texttt{Height} \tag{1}$$

   Allowing for variation among trees, and a more general relationship, we might consider the model

$$\texttt{Volume} = \gamma \cdot (\texttt{Girth})^{\beta_1} \cdot (\texttt{Height})^{\beta_2} \cdot e \tag{2}$$

   where e is multiplicative error.

   (a) [2 pts] Use transformations to *linearize* the model (2), i.e. transform it to have the form of a linear regression model in some transformed variables.

   (b) [2 pts] Fit the linearized model, and produce a summary of your results.

[1] Data set `trees` is actually in the `datasets` package, which is automatically available, so you do not need any other package to load it.

(c) [2 pts] Briefly assess the fit of your model using diagnostic plots.

(d) [2 pts] Form individual 95% confidence intervals for β_1 and β_2. Do they contain the theoretical values suggested by the relationship (1)?

(e) [2 pts] Consider a new tree (of the same kind) that has a girth of 10.9 inches and a height of 75 feet. Using the fitted model, form a 95% *prediction* interval for its *log*-volume.

(f) [2 pts] Transform your prediction interval (from the previous part) back to the original volume scale (in cubic feet).

3. Use the `ais` data set (in package `alr4`) with `Bfat` as the response and *only* the variables `Sex`, `Ht`, `Wt`, `LBM`, `BMI`, and `SSF` as possible predictors. Implement the following variable selection methods to determine a model. In each case, (i) *show* appropriate R output, and (ii) *list* the independent variables in the final model.

   (a) [2 pts] forward selection (use $F_{in} = 3$)

   (b) [2 pts] backward elimination (use $F_{out} = 3$)

   (c) [2 pts] selection with the R function `leaps`, according to minimum C_p
   (Hint: When you use `model.matrix`, use the formula `Bfat ~ Sex + Ht + Wt + LBM + BMI + SSF - 1`.)

   (d) [2 pts] stepwise selection with the R function `step`, using the arguments `object= lm(Bfat ~ 1, data=ais)`, `scope= ~ Sex + Ht + Wt + LBM + BMI + SSF`, and `direction= "both"`. (Consult `help(step)` for more information.)
   (Note: This will perform stepwise selection using AIC as the selection criterion.)

4. [GRADUATE SECTION ONLY] The so-called *arcsine transformation*, often used to transform binomial proportions, is given by

$$h(y) \; = \; \sin^{-1}(\sqrt{y}), \qquad 0 \le y \le 1$$

   (a) [2 pts] Compute the first derivative of $h(y)$ (for $0 < y < 1$).

   (b) [2 pts] Using the method demonstrated during lecture, verify that the arcsine transformation is (approximately) variance-stabilizing for the situation

$$\mathrm{Var}(Y) \; \propto \; \mathrm{E}(Y)\left(1 - \mathrm{E}(Y)\right)$$

   (c) [4 pts] $Y = W/n$ is a *binomial proportion* if $W \sim \mathrm{binomial}(n, p)$. Derive the mean and variance of Y (using the mean and variance of W, which you know).

   (d) [2 pts] Briefly explain why binomial proportions (with the same n) satisfy the condition of part (b).

Some reminders:

- Unless otherwise stated, all data sets are either automatically available or can be found in either the `alr4` package or the `faraway` package in R.

- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.